Atmospheric Environment 79 (2013) 822–830

Contents lists available at ScienceDirect

Atmospheric Environment

journal homepage: www.elsevier.com/locate/atmosenv

Air quality prediction using optimal neural networks with stochastic variables

Ana Russo [a,*], Frank Raischel [b], Pedro G. Lind [b,c,d]

[a] Center for Geophysics, IDL, University of Lisbon, 1749-016 Lisboa, Portugal
[b] Center for Theoretical and Computational Physics, University of Lisbon, Av. Prof. Gama Pinto 2, 1649-003 Lisbon, Portugal
[c] TWIST – Turbulence, Wind Energy and Stochastics, Institute of Physics, Carl-von-Ossietzky University of Oldenburg, DE-26111 Oldenburg, Germany
[d] ForWind – Center for Wind Energy Research, Institute of Physics, Carl-von-Ossietzky University of Oldenburg, DE-26111 Oldenburg, Germany

HIGHLIGHTS

- Optimized variables selection on a multivariate system by stochastic data analysis.
- Selection of variables used as input for ANN air quality forecast models.
- Use of derived variables as input for ANN models maintains forecast capabilities.
- Use of derived variables as ANN's inputs reduces the amount of input variables.
- Methodology can be adapted to other ANN models in weather or geophysical forecast.

ARTICLE INFO

Article history:
Received 18 March 2013
Received in revised form
26 July 2013
Accepted 29 July 2013

PACS:
[2010] 92.60.Sz
02.50.Ga
02.50.Ey

Keywords:
Pollutants
Neural networks
Stochastic systems
Environmental research

ABSTRACT

We apply recent methods in stochastic data analysis for discovering a set of few stochastic variables that represent the relevant information on a multivariate stochastic system, used as input for artificial neural network models for air quality forecast. We show that using these derived variables as input variables for training the neural networks it is possible to significantly reduce the amount of input variables necessary for the neural network model, without considerably changing the predictive power of the model. The reduced set of variables including these derived variables is therefore proposed as an optimal variable set for training neural network models in forecasting geophysical and weather properties. Finally, we briefly discuss other possible applications of such optimized neural network models.

1. Introduction

Urban air pollution is a complex mixture of toxic components with considerable impact on the inhabitants of urban regions, particularly those belonging to sensitive groups, such as children and people with previous heart and respiratory insufficiency (Kolehmainen et al., 2001). Therefore, forecasting the temporal evolution of air pollution concentrations in specific urban locations emerges as a priority for guaranteeing life quality in urban and metropolitan centers. With this aim, and in order to identify and predict in advance episodes of low air quality at regional and local scales, air quality forecasting models have been developed, considering the characteristics of atmospheric pollution and its consequent impact on people's health and life quality.

Straightforward approaches such as Box models (Middleton, 1997), Gaussian plume models (Reich et al., 1999), persistence and regression models (Shi and Harrison, 1999) are commonly applied to characterize and forecast air pollutants dispersion. These models are easy to implement and allow for the rapid calculation of forecasts. However, they include significant simplifications (Luecken et al., 2006) and usually do not describe the processes and interactions that control the transport and chemical

* Corresponding author.
E-mail addresses: ana_russo@yahoo.com, acrusso@fc.ul.pt (A. Russo).

1352-2310/$ – see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.atmosenv.2013.07.072

behavior of pollutants in the atmosphere (Luecken et al., 2006), important for instance for secondary pollutants (Sokhi et al., 2006). Improvements have been made with deterministic dispersion models and statistical-based approaches, which however, being highly nonlinear (Lal and Tripathy, 2012), require a large amount of accurate input data and are considerably expensive from the computational point of view (Dutot et al., 2007).

A promising alternative to all these models are artificial neural networks (ANN) (Lal and Tripathy, 2012; Nejadkoorki and Baroutian, 2012; Gardner and Dorling, 1998). Several ANN models have already been used for air quality forecast, in particular for forecasting hourly averages (Kolehmainen et al., 2001; Perez et al., 2000; Kukkonen et al., 2003) and daily maxima (Perez, 2001). Further, several authors compared already the potential of different approaches when applied to different pollutants and prediction time lags (Kukkonen et al., 2003; Yi and Prybutok, 2002; Gardner and Dorling, 2000; Hooyberghs et al., 2005). Still, though successful in many situations and having considerably less restrictions on the input data, large training data sets are usually required to improve accuracy and minimize uncertainty in the output data, which up to now has been a significant disadvantage of these models.

Recently, we applied methods from stochastic data analysis and statistical physics for deriving variables with reduced stochastic fluctuations (Vasconcelos et al., 2011) to empirical data in sets of NO_2 concentration measurements (Raischel et al, 2012). Such methods were introduced in the late nineties (Friedrich and Peinke, 1997; Friedrich et al., 2011) for analyzing measurements on complex stochastic processes, aiming for a quantitative estimation of drift and diffusion functions from sets of measurements that fully define the evolution equation of the underlying stochastic variables. The framework has already been applied successfully, for instance to describe turbulent flows (Friedrich and Peinke, 1997) and the evolution of climate indices (Lind et al., 2005, 2007), performance curves of wind turbines (Raischel et al, 2013), stock market indices (Friedrich et al., 2000), and oil prices (Ghasemi et al., 2007). At the same time, the basic method has been refined in particular for data with low sampling frequency (Kleinhans et al., 2005; Lade, 2009) and subjected to strong measurement noise (Boettcher et al., 2006; Lind et al., 2010; Carvalho et al., 2011).

In this paper we present an important application of such variables: using them as input for training ANN enables one to reduce considerably the amount of input data needed for achieving a given accuracy. We argue that this reduction in the number of input variables is possible because the derived variables incorporate temporal correlations between independent and spatially separated monitoring stations. Moreover, as we quantitatively show below, when using this reduced amount of information that includes the derived variables, the predictive power of the ANN is not significantly changed, which is a major advantage when working with observational data which might include missing values. Combining a faster ANN training with the same predictive power may improve the ability and capability of alert system for air quality in large urban centers. We start in Sec. 2 by briefly describing ANN models as well as the main points of the stochastic data analysis procedure used. In Sec. 3 the empirical data is described, comprising two different data sets of NO_2 concentration measures in the city of Lisbon, Portugal (see Fig. 1). In Sec. 4 the results are discussed in the light of predictive power measures, and Sec. 5 concludes the paper.

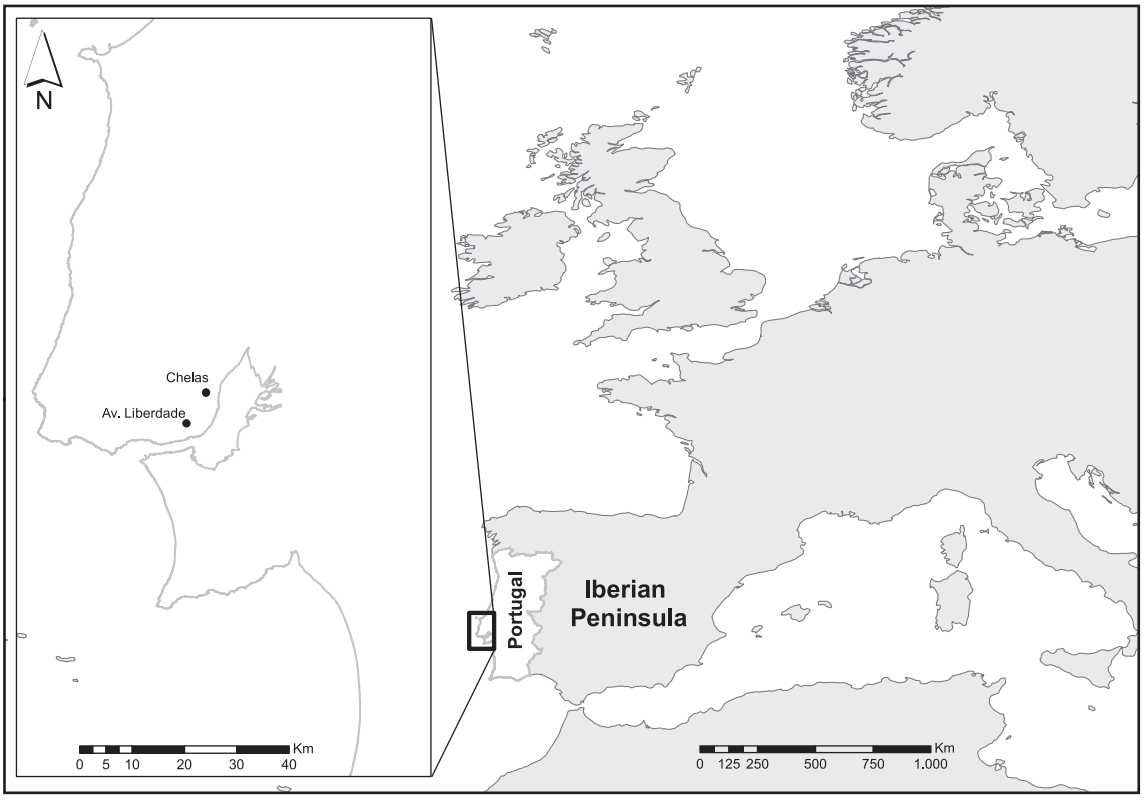

Fig. 1. NO_2 measurement stations in the region of Lisbon (Portugal) at the Southwestern coast of Europe. In this paper we focus on the set of measurements taken at the stations of Chelas and Avenida da Liberdade, approximately 4 km distant apart, with approximately $3x10^4$ data points. Each data set is extracted within the period between 2002 and 2006, with a frequency of 1 h^{-1}.

2. Methods

2.1. The neural network framework

Artificial neural network models are mathematical models inspired by the functioning of nervous systems (Gardner and Dorling, 1998; Cobourn et al., 2000; Agirre-Basurko et al., 2006), which are composed by a number of interconnected entities, the artificial neurons (see Fig. 2).

These neurons may be associated in many different ways (Agirre-Basurko et al., 2006; Haykin, 1999), depending on the characteristics of the proposed problem. To construct an ANN model the air pollution system is considered as a system that receives information from n distinct sets of inputs X_i ($i = 1,\ldots,n$), namely weather parameters and air pollution properties, and produces a specific output, in our case the concentration of the NO_2 pollutant (Gardner and Dorling, 1998). No prior knowledge about the relationship between input and output variables is assumed. The input variables should be independent from each other and each one is represented by its own input neuron $i = 1,\ldots,n$. Each neuron computes a linear combination of the weighted inputs ω_{ij}, including a bias term b_i, from the links feeding into it and the corresponding summed value $C_j = \sum_i \omega_{ij} X_i + b_i$ is transformed using a function f, either linear or non-linear such as log-sigmoid or hyperbolic tangent. 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 (e.g. dummy input with a magnitude equal to 1). The output obtained is then passed as a new input $\tilde{X}_j = f(C_j)$ to other nodes in the following layer, usually named hidden layer. Though one is allowed to use several neurons in this hidden layer, it is generally advantageous to somehow minimize the number of hidden neurons, in order to improve the generalization capabilities of the model and also to avoid over-fitting. In particular, a simple one-layer ANN structure with just one neuron employing a linear activation function reduces to the well-known linear regression model (Weisberg, 1985). The ANN models used here are based on a feed-forward configuration of the multilayer perceptron that has been used by several authors (Hooyberghs et al., 2005; Papanastasiou et al., 2007). A large number of architectures has been tested. The use of two layers was verified to be sufficient. The use of more layers was concluded to be redundant, and therefore we only use two layers, one input layer and one hidden layer.

Having such a framework of input variables and sets of functions, the ANN has to be trained in order to obtain the best estimate for each weight ω. The weight values are determined by an optimization procedure, the so-called learning algorithm (Haykin, 1999), which in our case uses a cross-validation procedure (Wilks, 2006) to ensure stability of the model. The cross-validation was applied dividing the available period into four sets and completing the calibration–validation procedure four times independently, i.e., from the 4 years of data available, data from 3 years were used to build the model and data from 1 year for validation. The validation year was then cycled through the 4 year period. For each validation year, a set of performance measures were computed between the observed real values and the ANN forecasts, namely, the Pearson correlation coefficient (PC); the root mean square error (RMSE); and the skill against persistence (Skillp). The outcomes of each resulting performance measure were analyzed for each year and then averaged for the complete period, resulting in an average value for each monitoring station.

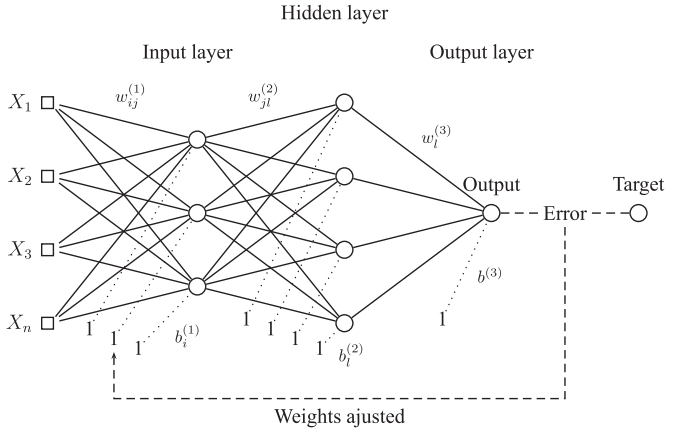

Fig. 2. Example of the structure for a feed-forward artificial neural network model with three layers. Variables $X_i = {}_{1,\ldots,n}$ represent the input variables (neurons), ω the weights associated to each neuron and b the bias vectors which combined will produce an output within certain error limits.

There are several learning algorithms, depending on whether the ANN model is linear or non-linear. For linear ANN models, the learning algorithm is typically based on the Widrow–Hoff learning-rule, also known as the least mean square rule, which produces a unique solution corresponding to the absolute minimum value of the error surface (Trigo and Palutikof, 1999). For non-linear models, the back-propagation (BP) is one of the most popular and common training procedures used, which is described in depth in the literature (Haykin, 1999; Trigo and Palutikof, 1999). It has been shown in literature that BP training algorithms have two caveats: convergence may be slow and the final weights may be trapped in local minima over the highly complex error surface (Trigo and Palutikof, 1999). As an alternative, the Levenberg–Marquardt method (Press et al., 1992) minimizes an error function in "damped" procedures, i.e. selecting steps proportional to the gradient of the error function. The Levenberg–Marquardt method requires more memory (Haykin, 1999; Trigo and Palutikof, 1999) than the Widrow–Hoff rule, but has the advantage of converging faster and with a higher effective robustness than most BP schemes (Trigo and Palutikof, 1999) because it avoids having to compute second-order derivatives. The Levenberg–Marquardt method was applied for this study. Together with the 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, i.e., the forecast for each time step simply corresponds to the value of the previous time step, the linear regression model constitutes the baseline against which the performance of non-linear ANN models are usually compared. Due to a certain level of memory that characterizes air pollutants, persistence corresponds to a benchmark model considerably more difficult to beat than climatology (Demuzere et al., 2009).

2.2. Deriving optimal stochastic variables

In this section we briefly describe how from a number K of sets of measurements, one is able to derive a set of few stochastic variables containing information from all of them.

This procedure assumes that corresponding to each measurement variable there is a property that evolves according to some stochastic equation. More precisely, the K-dimensional state vector $X = (x_1,\ldots,x_K)$ characterizes the set of K measurements at each time-step and evolves according to the Itô–Langevin equations (Risken, 1984; Gardiner, 1997):

$$\frac{dX}{dt} = h(X) + g(X)\Gamma(t), \tag{1}$$

where $\Gamma = (\Gamma_1,\ldots,\Gamma_K)$ is a set of K independent stochastic forces with Gaussian distribution fulfilling $\langle \Gamma_i(t) \rangle = 0$ and

$\langle \Gamma_i(t)\Gamma_j(t')\rangle = 2\delta_{ij}\delta(t-t')$. On the right hand side of Eq. (1) the term with function $h = \{h_i\}$ describes the deterministic part, which drifts the system, while the term with $g = \{g_{ij}\}$ accounts for the amplitude of the stochastic contributions characterized through the properties of Γ (Friedrich et al., 2011).

The heart of the method lies in the fact that the coefficients h and g are closely related to the drift vectors and diffusion matrices describing the evolution of the joint probability density function of the vector state X by means of the corresponding Fokker–Planck equation (Risken, 1984; Gardiner, 1997). As has been shown previously in other contexts (Friedrich et al., 2011, 2000; Lind et al., 2005; Ghasemi et al., 2007; Kleinhans et al., 2005; Boettcher et al., 2006; Lind et al., 2010), the drift vector and the diffusion matrix can therefore be extracted directly from the data set Y (Raischel et al, 2012)

$$D_i^{(1)}(X) = h_i(X) \approx \lim_{\tau\to 0}\frac{1}{\tau}\langle Y_i(t+\tau) - Y_i(t) | Y(t) = X\rangle \tag{2a}$$

$$D_{ij}^{(2)}(X) = \sum_{k=1}^{K} g_{ik}(X)g_{jk}(X) \approx \lim_{\tau\to 0}\frac{1}{2\tau}\langle (Y_i(t+\tau) - Y_i(t)) \times (Y_j(t+\tau) - Y_j(t)) | Y(t) = X\rangle \tag{2b}$$

for $i,j = 1,\ldots,K$ and where $\langle\cdot|Y(t) = X\rangle$ symbolizes conditional averaging overall measurements $Y(t)$ that fulfill the condition $Y(t) = X$.

The last equation in both Eqs. (2a) and (2b) yields the operational definition of the first and second conditional moments (Friedrich et al., 2011; Lind et al., 2010), respectively. The limit in Eq. (2) is typically approximated by the slope of a linear fit of the corresponding conditional moments at small τ. When this linear fit is not possible, as it is the case for the NO_2 measurements under consideration, an alternative estimate (Kleinhans et al., 2005) is to consider the first value of $M(\tau)/\tau$ at the lowest value of τ. Additional analysis has been carried out (Raischel et al, 2012), namely confirming the Markovian properties of the data sets, which is a prerequisite for assuming a Langevin process, Eq. (1). In case Markovian properties are not observed, it should be noted that this method may be still applied with alternative procedures (Boettcher et al., 2006; Lind et al., 2010; Carvalho et al., 2011). Furthermore, in case a Fourier spectrum shows periodicities, those should be filtered out by a proper detrending procedure. More details about how to apply Eq. (2) can be found in Raischel et al (2012).

We apply this framework to the two-dimensional system of detrended NO_2 concentration measurements taken in two stations in Lisbon situated in Chelas and Avenida da Liberdade. We consider therefore a vector $X = (x_1,x_2)$. Both sets of measurements do exhibit Markovian properties (Raischel et al, 2012), and therefore both drift and diffusion functions are properly derived.

To arrive to the optimal variables, we next determine the eigensystem of the diffusion matrix and investigate its principal directions (Vasconcelos et al., 2011; Gradišek et al., 2003; van Mourik et al., 2006). These principal directions are computed for each mesh point previously defined in phase space (Raischel et al, 2012). The aim is to obtain the transform of the original coordinates $X = \{x_i\}$ into new ones $\tilde{X} = \{\tilde{x}_i\}$, such that the diffusion matrix is diagonalized. The transformation $\tilde{X} = F(X,t)$ is a two-times continuously differentiable function. In general, the eigenvalues of the diffusion matrix indicate the amplitude of the stochastic force and the corresponding eigenvector indicates the phase space direction towards which such force acts.

In this way, the stochastic contribution is decoupled for each new variable. Consequently, if the eigenvalues in the transformed coordinates are significantly different, we are able to restrict our investigation to the coordinates with lower stochastic sources, i.e. lower eigenvalues. For such eigenvalues, the vector field of their eigenvectors defines the path in phase space towards which the fluctuations are minimal. Additionally, if these j eigenvalues are very small compared to all the others, the corresponding stochastic forces can be neglected and the system can be assumed to have only $K–j$ independent stochastic forces, reducing the number of stochastic variables in the system. For all the details see Raischel et al (2012).

3. Data

3.1. Target data

We consider hourly measurements of NO_2 concentrations in the metropolitan region of Lisbon, Portugal, namely at the monitoring stations of Chelas (C) and Avenida da Liberdade (AL) (see Fig. 1). The data were recorded from 2002 to 2006, corresponding to $\sim 3x10^4$ measurement points with roughly 1% of discarded values, due to incomplete or erroneous measurements. The stations are located at a distance of 4.8 km from each other. In the following, the NO_2 concentrations at the stations of Chelas and Avenida da Liberdade will be designated as $Y_C(t)$ and $Y_{AL}(t)$, respectively, omitting the temporal dependency when not necessary. Fig. 3 shows the Fourier spectrum of both these data series.

3.2. Input data for ANN training

The ANN input data sets consist of the aforementioned hourly NO_2 concentration measurements, and of concentrations of two other pollutants, namely NO and CO, also measured at the monitoring stations of Chelas and Avenida da Liberdade, from January 1st 2002 until December 31st of 2006.

The first four years are used to construct the models and year 2006 is used for independent evaluation. More specifically the prediction is done in two steps. In the first step, we consider only the period 2002–2005. For this four years we take the first three, 2002–2004, for training the ANN and derive the respective parameter values of the ANN model. Using that ANN we predict

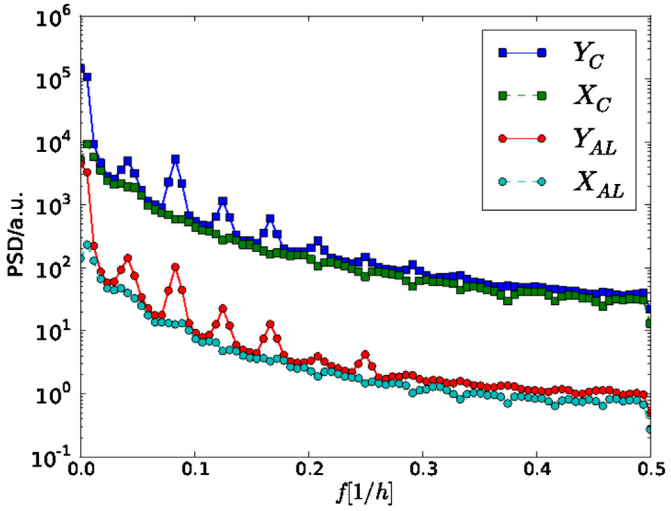

Fig. 3. Fourier spectrum (power spectral density PSD) of the NO_2 measurements in Chelas and Avenida da Liberdade stations. While the original measurements, symbolized by Y_C and Y_{AL}, show some periodic modes resulting from the daily, three-day and weekly cycles (see text), the detrended variables, X_C and X_{AL}, both in Chelas and Avenida da Liberdade have these periodicities suppressed. Plots are vertically offset for better visibility.

year 2005. Then we consider 2002, 2003 and 2005 for training the ANN, and predict 2004. Similar procedure is done for predicting 2002 and 2003. In the second step, we use the parameter values obtained in the first step, namely, weights and biases, for predicting 2006.

Besides pollutant concentrations, we also consider daily maximum temperature, daily mean wind direction and speed, daily humidity, daily radiance, hourly mean temperature, hourly pressure and hourly relative humidity, boundary layer height (BLH) from the European Center for Medium Weather Forecast (ECMWF), circulation weather type (CWT) at the regional scale determined for Portugal according to Trigo and DaCamara (Trigo and DaCamara, 2000), North Atlantic Oscillation (NAO) index from NCEP-NOAA, two weekly cycles and two yearly cycles. Specifically, BLH fields were retrieved from the 3 hourly ECMWF 40 years reanalysis (http://data-portal.ecmwf.int/data) for the 2002–2006 period. Afterward, we extracted the 00:00 UTC (BLH1), 03:00 UTC (BLH2), 9:00 (BLH3) and 21:00 UTC (BLH4) data from the retrieved BLH fields. Together with these variables we consider the two transformed variables derived through the procedure described in Sec. 2.2, before we suggest a method for reducing the number of input variables. In total there are 48 variables that are available as input data for the ANN model (See Table 1).

The two transformed variables are obtained from the original NO_2 measurements Y_C and Y_{AL} by first detrending them. As shown in Fig. 3 both sets of measurements Y_C and Y_{AL} have periodic contributions, which must be filtered out. These periodicities describe daily, weekly, seasonal and yearly variations of the concentration due to anthropogenic routines and to periodic atmospheric processes (Kolehmainen et al., 2001). In particular, the 24 h and one week cycles are both traffic related and mirror daily and weekly cycles. By filtering out these cycles, through a proper detrending, one is reduced to the stochastic contribution solely, which can be modeled through a stochastic differential equation (Raischel et al, 2012; Risken, 1984), having both the deterministic forcing in the drift vector h, and the diffusive fluctuation, g.

The detrended series derived from the original measurements Y_C and Y_{AL}, represented henceforth as X_C and X_{AL}, respectively, are obtained as follows. The data is partitioned into segments of length N, multiple of all relevant periodic modes. Our simulations have shown that two such partitions are needed in the present case, as one partition alone is not sufficient to remove the periodicities completely. First, averages over $N = 52$ weeks are performed, and afterward a second detrending with $N = 1$ day follows on consecutive periods of 14 days. Next, a mean segment is calculated by averaging measurements with the same relative position in the periodic segment. Finally, the detrended data set is obtained by subtracting the respective values of the mean segment from the measured data (Raischel et al, 2012). Whereas the Markovian framework can strictly only be applied after removing all the periodicities by detrending (or a similar procedure), we have verified that the estimates of the drift and diffusion coefficients and of the mean orientation angle are not altered significantly by the second step of the detrending procedure, *i.* e. after removing the periodicities only partially, which confirms that our detrending method does not introduce artifacts.

Table 1
Description of available input parameters for prediction. Here the time-lags are $\Delta t = 1,3,6,12$ and 24 h. For model M_s one considers a total maximum number of 46 variables and for model M_{s+} the two transformed variables V_0^+ and V_0^- are added.

Model	Variables	Time t and lag Δt
M_s	NO_2	$t-\Delta t$ to $t-24$
	NO, CO	$t-\Delta t$
	NAO index	t, $t-\Delta t$
	CWT	t, $t-1$
	BLH1, BLH2, BLH3, BLH4	$t-\Delta t$
	Daily maximum temperature	$t-\Delta t$
	Daily mean wind direction	$t-\Delta t$
	Daily mean wind speed	$t-\Delta t$
	Daily mean humidity	$t-\Delta t$
	Daily mean radiance	$t-\Delta t$
	Hourly mean temperature	$t-\Delta t$
	Hourly mean pressure	$t-\Delta t$
	Hourly mean relative humidity	$t-\Delta t$
	sin $(2\pi t/365)$, cos $(2\pi t/365)$	t
	sin $(2\pi t/7)$, cos $(2\pi t/7)$	t
M_{s+}	The same as M_s	(See above)
	V_0^+, V_0^-	$t-1$

As indicated by Fig. 3, the detrended data do not show patterns of periodicity, satisfying both Markov properties as well as the delta-correlated signature of the Gaussian-like noise (Raischel et al, 2012). One therefore may consider the series X_C and X_{AL} as a set described by two coupled Langevin Equations.

Having obtained the detrended data, the method described in Sec. 2.2 is then applied, yielding a two-dimensional drift vector and diffusion matrix. The diffusion matrix is diagonalized, yielding two eigenvalues and their normalized eigenvectors, orthogonal to each other. Multiplying them by their corresponding eigenvalues yields two orthogonal vectors that define an ellipse in phase space (X_C,X_{AL}). These diffusion ellipses have a specific orientation defined through an angle whose absolute value quantifies the relative off-diagonal contribution that describes the coupling of the noise terms by the diffusion matrix (Raischel et al, 2012). Rotating all the ellipses by the orientation angle, aligns the largest eigenvector along one of the coordinate axis and the smallest eigenvector along the other one. At each mesh point we construct from the numerical values of the two detrended time series, and the respective functional dependency of the eigenvalues on the original variables, two transformed time series, V_0^+ and V_0^-, which we take as additional input data for the ANN model. We then test the resulting improvement of the forecasts. The optimal variable corresponding to the largest eigenvalue is henceforth symbolized by V_0^+ while the other will be represented by V_0^-. As reported below, the variable V_0^+ shows a higher rank of importance than the variable V_0^- when selecting the most important variables for ANN training.

To select a reduced set of the M_{s+} input variables we conducted a forward stepwise regression (FSR) between the meteorological and the air quality variables for each monitoring station independently. The FSR allows variables to be optimized in order to predict NO_2 at each monitoring station. This procedure starts with the variable most correlated with the target, and adds one new variable which, together with the previous one, most accurately predicts the target, i.e. the variable that reduces the predicting error the most. One adds iteratively new variables in this way, ordering the set of M_{s+} variables by descending order of correlation with the target. The procedure stops when any new variable does not significantly reduce the prediction error. The corresponding significance of error reduction is measured by a partial F-test (Press et al., 1992). With this approach, given an initial set of variables for one monitoring station, one is able to select the best subset of variables for predicting the evolution of NO_2 concentration at that station.

In this study, we consider four distinct sets of variables for the input data, defining therefore four distinct ANN models.

The first set is the complete set of (up to, depending on the lag investigated) initial variables, neglecting the transformed variables V_0^+ and V_0^-, and is used for training the ANN, which then defines the standard ANN model, henceforth represented by M_s. Another set corresponds to the first set with the transformed variables V_0^+ and V_0^-, yielding the model M_{s+}.

The third set is extracted from the M_{s+} variables by FSR, yielding the model $M_{s,FSR}$.

The forth set is a subset of this extended set and uses the correlation ordering made for all the variables, where it considers only the variable V_O^+ and the variables having stronger correlation with the target. This is our optimal model M_O.

Table 2 shows the number of variables used for each model and for five different time lags. Depending on this time lag, namely $t-1$, $t-3$, $t-6$, $t-12$ and $t-24$ h, the input data includes hourly data from the previous $t = 1,\ldots,24$ h, $t = 3,\ldots,24$ h, $t = 6,\ldots,24$ h, $t = 12,\ldots,24$ h and $t = 24$ h, respectively.

While model M_{s+} uses always two more variables than the standard model M_s, the number of variables for the optimal model M_O is typically smaller then the number of variables for $M_{s,\mathrm{FSR}}$.

In the last column of Table 2, we indicate the relative reduction on the number of input variables when using model M_O instead of model $M_{s,\mathrm{FSR}}$. The reduction of the number of variables used for training the ANN model is typically above 50%, with only three exceptions, raising up to 90% for the largest time-lag predictions. Notice that by construction the number in column M_O indicates the rank of the variable V_O^+ when ordering the variables according to their correlation with the target. This result also shows the strong correlation between our transformed variable V_O^+ and the target.

In all cases, the variable V_O^+ shows a higher correlation with the target than the variable V_O^-, therefore V_O^- does not appear in the subsets for M_O. The reason for this higher correlation is associated with the strength of the corresponding fluctuations. Since the variable V_O^+ is the transformed variable corresponding to the largest eigenvalue of the diffusion matrix for the coupled system of both NO_2 concentrations, X_C and X_{AL}, the pair of variables fluctuate stronger along this direction and therefore contains a larger part of the correlations than the other transformed variable V_O^-. This situation is somehow comparable to the assumptions of the principal component analysis model (Pearson, 1901).

4. Stochastic variables as optimal input for neural networks

In the previous section we concluded that one of our transformed variables is more correlated with the target than most of the meteorological and air quality variables. Collecting only variables with an equal or higher correlation than the one observed for V_O^+, one obtains the model M_O, which must now be compared with the standard model $M_{s,\mathrm{FSR}}$ in its predictive power. In this section we present such comparison between both models. Using different measures of predictive power, we conclude that, in general, the optimal model most frequently evidences the same predictive power as model $M_{s,\mathrm{FSR}}$. Results are summarized in Table 3.

The overall conclusion is that, despite the significant decrease of input data, the predictive power of model M_O is not worse than the one of the models $M_{s,\mathrm{FSR}}$.

Table 2
Number of retained variables for each model at both *C* and *AL* monitoring stations. Five different time lags Δt are considered. The last column indicates the relative reduction of the number of input variables when substituting the standard model $M_{s,\mathrm{FSR}}$ by model M_O.

	Δt (h)	M_s	M_{s+}	$M_{s,\mathrm{FSR}}$	M_O	Dev.
C	1	46	48	22	13	41%
	3	44	46	26	26	0%
	6	41	43	31	15	55%
	12	36	38	28	7	75%
	24	27	29	19	2	89%
AL	1	46	48	29	13	52%
	3	44	46	25	18	28%
	6	41	43	28	13	54%
	12	36	38	27	10	63%
	24	27	29	20	2	90%

To evaluate the efficiency and performance of our method four other quantities are computed. The first such quantity is the well-known Pearson correlation coefficient ("PC") (Pearson, 1901),

$$\mathrm{PC} = \frac{\sum_{i=1}^{N}(y_i-\overline{y})(o_i-\overline{o})}{\left[\sum_{i=1}^{N}(y_i-\overline{y})^2\sum_{i=1}^{N}(o_i-\overline{o})^2\right]^{1/2}}, \tag{3}$$

where y_i denotes the respective model forecast at time i and o_i denotes the real observed values at time i.

From Table 3 one sees that the difference in the Pearson correlation between both models, $M_{s,\mathrm{FSR}}$ and M_O, is almost nonexistent, except for the largest time-lag, namely $t-24$. The PC results for the independent sample (2006) are relatively lower than for the calibration–validation period as expected, with the exception of the PC for AL ($t = 24$ h).

The performance of both our method and the standard one indicates a raise in PC for the 24-hour lag because of the daily periodicity.

In Table 3 we also note two outliers in the PC and the Skillp for the % Dev for the 24 h time lag at the Chelas station in the independent validation period. This has two reasons. First, the PC of $M_{s,\mathrm{FSR}}$ in Chelas is very low (45.8) compared to the Av. Liberdade station (61.1), due to a large amount of missing measurement values during 2006. Our optimized model M_o is more robust than

Table 3
Comparative overview of the efficiency and performance for ANN according to two of the models described in Sec. 3.2. To evaluate the efficiency and performance we use the Pearson coefficient ("PC"),Eq. (3), the skill against persistence ("Skillp"), Eq. (4), the root mean square error ("RMSE"),Eq. (5), and the mutual information ("MI"), Eq. (6). The deviations ("Dev") give the relative improvement (or deterioration) of the optimized model M_O vs. $M_{s,\mathrm{FSR}}$. As an overall conclusion, one can state that the inclusion of the stochastic variable produces similar results than the standard model, using much less input variables. In parenthesis one shows the result obtained through independent validation (see text). The 24 h % Dev at the Chelas station for PC and Skillp in the validation period (+23.1 and −25.9) are outliers due to insufficient measurement data (see text). (*) Not sufficient data to forecast.

Lag (h)	Sta.	Mod.	PC (%)	Skillp (%)	RMSE ($\mu g\ m^{-3}$)	MI
1	C	$M_{s,\mathrm{FSR}}$	92.8 (92.0)	16.8 (54.5)	12.9 (8.9)	0.8 (0.8)
		M_O	92.8 (91.8)	16.1 (31.2)	13.0 (8.7)	0.8 (0.8)
		% Dev	0 (−0.2)	−4.1 (−42.8)	+0.8 (−2.2)	0 (0)
	AL	$M_{s,\mathrm{FSR}}$	92.8 (92.0)	17.2 (48.9)	12.8 (16.8)	0.8 (0.6)
		M_O	92.8 (92.0)	16.5 (46.1)	12.9 (16.7)	0.8 (0.5)
		% Dev	0 (0)	−4.1 (−5.7)	+0.8 (−0.6)	0 (−16.7)
3	C	$M_{s,\mathrm{FSR}}$	76.8 (*)	33.2 (*)	22.2 (*)	0.5 (*)
		M_O	76.8 (67.2)	33.2 (52.5)	22.2 (25.9)	0.5 (0.3)
		% Dev	0 (*)	0 (*)	0 (*)	0 (*)
	AL	$M_{s,\mathrm{FSR}}$	76.9 (71.8)	33.9 (56.3)	22.0 (30.6)	0.5 (0.3)
		M_O	76.8 (71.2)	33.8 (53.8)	22.0 (30.8)	0.4 (0.3)
		% Dev	−0.1 (−0.8)	−0.3 (−4.4)	0 (+0.7)	−20 (0)
6	C	$M_{s,\mathrm{FSR}}$	65.2 (44.8)	47.1 (65.5)	26.3 (20.2)	0.3 (0.2)
		M_O	65.0 (46.0)	47.0 (59.5)	26.3 (19.2)	0.3 (0.2)
		% Dev	−0.3 (+2.7)	−0.2 (−9.1)	0 (−5.0)	0 (0)
	AL	$M_{s,\mathrm{FSR}}$	65.4 (57.4)	47.8 (67.3)	26.0 (35.6)	0.3 (0.2)
		M_O	65.2 (57.8)	47.5 (66.9)	26.1 (35.1)	0.3 (0.2)
		% Dev	−0.3 (+0.7)	−0.6 (−0.6)	+0.4 (−1.4)	0 (0)
12	C	$M_{s,\mathrm{FSR}}$	62.2 (39.0)	49.1 (60.4)	27.2 (20.9)	0.3 (0.1)
		M_O	61.7 (38.4)	48.6 (60.0)	27.3 (20.9)	0.3 (0.1)
		% Dev	−0.8 (−1.5)	−1.0 (−0.7)	+0.4 (0)	0 (0)
	AL	$M_{s,\mathrm{FSR}}$	63.3 (50.4)	50.4 (66.1)	26.6 (37.9)	0.3 (0.2)
		M_O	62.7 (52.0)	49.8 (66.1)	26.8 (37.3)	0.3 (0.2)
		% Dev	−0.9 (+3.2)	−1.2 (0)	+0.8 (−1.6)	0 (0)
24	C	$M_{s,\mathrm{FSR}}$	60.1 (45.8)	33.9 (54.9)	27.7 (19.7)	0.2 (0.2)
		M_O	56.8 (56.4)	29.9 (40.7)	28.5 (18.1)	0.2 (0.2)
		% Dev	−5.4 (+23.1)	−11.8 (−25.9)	+2.9 (−8.1)	0 (0)
	AL	$M_{s,\mathrm{FSR}}$	59.9 (61.1)	33.8 (55.1)	27.6 (34.2)	0.2 (0.2)
		M_O	56.8 (62.7)	30.0 (55.3)	28.3 (33.3)	0.2 (0.2)
		% Dev	−5.2 (+2.6)	−11.2 (+0.4)	+2.5 (−2.6)	0 (0)

$M_{s,FSR}$, since the PC value (56.4) is much closer to the Av. Liberdade value (62.7). It therefore has a relatively much better PC than the standard model (+23.1%).

However, the time series are also strongly periodic with a 24 h period, which our optimized model with two variables cannot reproduce as well as the $M_{s,FSR}$ model. The corresponding Skillp value at the Chelas station is worse for the optimized model, by −25,9%. We verified numerically that excluding the invalid measurements removes these outliers.

Another important quantity is the skill against persistence ("Skillp") which can be interpreted as the percentage of improvement that our model can provide when compared with the persistence model, i.e. the forecast for a given hour is the observed value of the previous time lag, i.e., 1 h, 3 h, 6 h, 12 h or 24 h hours before. It is given by

$$\text{Skillp} = \frac{\frac{1}{N}\sum_{i=1}^{N}(y_i - o_i)^2 - \frac{1}{N-1}\sum_{i=1}^{N}(\hat{y}_i - o_i)^2}{\frac{1}{N-1}\sum_{i=1}^{N}(\hat{y}_i - o_i)^2}, \quad (4)$$

where $\hat{y}_i$ is the value of the time series variable y 1, 3, 6, 12 or 24 h before.

From Table 3 one sees that the decrease in the skill against persistence between both models, $M_{s,FSR}$ and M_O, is not significant ($\lesssim$4%) for short time-lags except for the largest time-lag, namely $t-24$. The skill against persistence results for the independent sample (2006) are higher than for the calibration–validation period for all the time lags.

The particular NO_2 data sets we used 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, with the occurrence of extreme weather events, namely an exceptional heatwave that struck the entire western Europe (2003), and one of the most severe droughts of the 20th century (2005). Air pollution is strongly influenced by shifts in the weather (e.g., heat waves or droughts). Thus, changes in the temperature, humidity, wind, and precipitation can have a deep impact on air quality because of induced changes in the transport, dispersion, and transformation of air pollutants at multiple scales (Dias et al., 2012). 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. The difference between independent and the calibration–validation samples tends to decrease with the time-lag, except for 24 h where periodicity plays again a role (see above).

Moreover checking again the PC column in Table 3, one observes a tendency for higher performance of the optimized model with independent validation, which is due to the mildness characteristics of year 2006, as we explained above.

Another quantity is the root mean square error ("RMSE") which represents the difference between the pairs of forecast and observation values and is given by

$$\text{RMSE} = \sqrt{\sum_{i=1}^{N}(y_i - o_i)^2}. \quad (5)$$

While for shorter time-lags there is almost no difference between both models, a slight increase ($\lesssim$3%) of RMSE is observed in the optimal model for the $t = 24$ time lag. The RMSE for the independent sample are usually lower for the M_O, except for the AL $t = 3$ case.

Finally, the mutual information evaluates the dependence between observations and predictions and is given by

$$\text{MI} = \sum_{y \in Y_{\text{bins}}} \sum_{o \in O_{\text{bins}}} p(y, o)\log\left(\frac{p(y, o)}{p(y)p(o)}\right) \quad (6)$$

where $p(y,o)$ is the joint probability of observations and predictions for the same time-steps and $p(y)$ and $p(o)$ are the corresponding marginal distributions, computed within the range of admissible values, Y_{bins} and O_{bins}, properly discretized. With one single exception in the 2002–2005 period and one in the 2006 period, no deviations are observed when comparing the mutual information in both models.

Enabling a significant reduction of input data - less than one half - and simultaneously presenting a predictive power at least as good as the one given by the standard model, one can conclude that the model M_O using our stochastic may be seen as a good alternative for establishing ANN models in air quality prediction. Fig. 4 shows for both Chelas and Avenida da Liberdade the scatter plot between observations (NO_2 measurements) and the corresponding predictions from an ANN model training with the optimized model M_O, indicating a consistently good agreement.

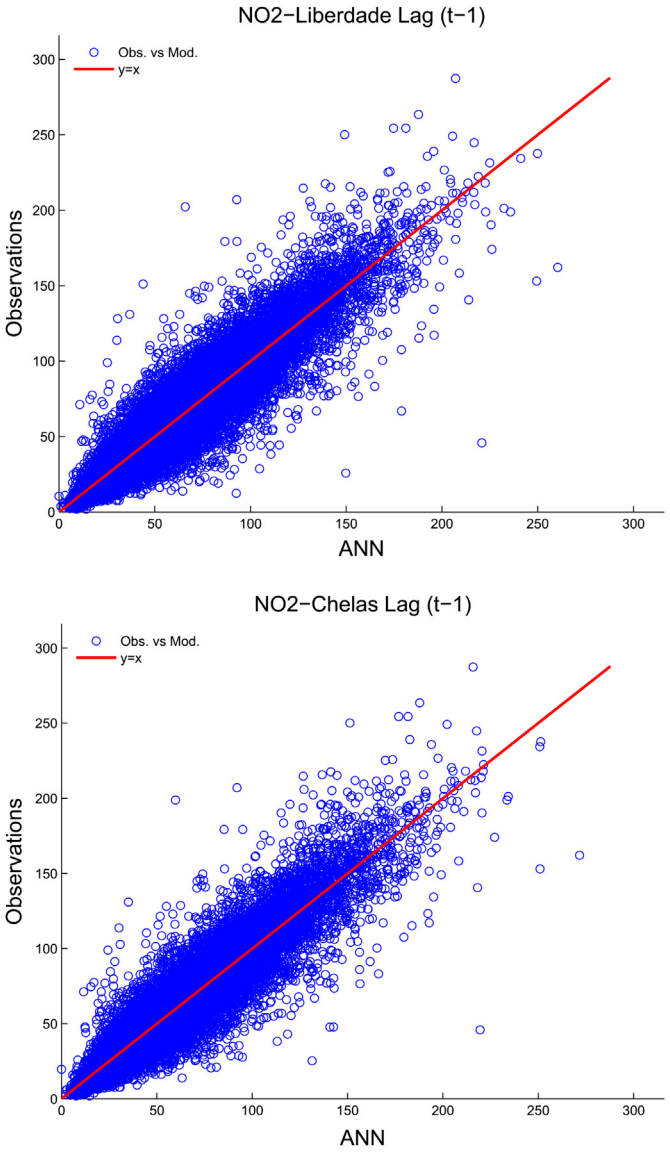

Fig. 4. Scatter plot between observations and predicted values from the ANN for each station separately. The ANN is trained by the input variables selected by forward stepwise regression (FSR) elaborated in Sec. 3.2. In all cases the time lag is 1 h.

One last remark should be added, concerning the nature and the data quality extracted at each one the two stations. While Chelas station is classified as a background station, with a more regular signal, the station in Av. Liberdade is an urban traffic station showing typically a largely fluctuating signal. Moreover, the particular series of Chelas due to particular local technical features has a larger number of missing values than the station at Av. Liberdade. This justifies to some extent the marked differences between the two locations especially with respect to the skill against persistence.

5. Conclusions

In this paper we applied a method for deriving eigenvariables in systems of coupled stochastic variables, which were then used as input variables to considerably reduce the amount of input data needed for training ANN models, typically by a factor of two and for large time-lags by a factor of ten. The predictive power is maintained. Indeed, the introduction of the stochastic variables as input data for training the ANN model allows to preserve the predictive power with considerable less input information.

This is because the variable incorporates temporal correlations between independent and spatially separated monitoring stations.

In the particular context of atmospheric environment, the reduction of the amount of input data optimizes the ANN models in the sense that it enables faster predictive outcome without changing significantly the predictive power. Thus, our stochastic variables together with the findings of this study can be taken as a first step for improving alarm systems, making them more efficient in predicting periods of lower levels in the air quality. Moreover, being a general numerical procedure for any given set of measurements, our finding can be easily adapted to other ANN 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 thank Ricardo Trigo for useful discussions and DAAD and FCT for financial support through the bilateral cooperation DREBM/DAAD/03/2009. FR (SFRH/BPD/65427/2009) and PGL (*Ciência 2007*) thank Fundação para a Ciência e a Tecnologia for financial support, also with the support Ref. PEst-OE/FIS/UI0618/2011. The authors would like to acknowledge Agência Portuguesa do Ambiente and European Centre for Medium Weather Forecast for providing the environmental and meteorological data, respectively.

References

Agirre-Basurko, E., Ibarra-Berastegi, G., Madariaga, I., 2006. Regression and multilayer perceptron-based models to forecast hourly O_3 and NO_2 levels in the Bilbao area. Environmental Modelling & Software 21, 430446.

Boettcher, F., Peinke, J., Kleinhans, D., Friedrich, R., Lind, P.G., Haase, M., 2006. Reconstruction of complex dynamical systems affected by strong measurement noise. Physical Review Letters 97, 090603.

Carvalho, J., Raischel, F., Haase, M., Lind, P.G., 2011. Evaluating strong measurement noise in data series with simulated annealing method. Journal of Physics 285, 012007.

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.

Demuzere, M., Trigo, R., Arellano, V., van Lipzig, N., 2009. The impact of weather and atmospheric circulation on O_3 and PM10 levels at a rural mid-latitude site. Atmospheric Chemistry and Physics 9, 26952714.

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.

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.

Friedrich, R., Peinke, J., 1997. Description of a turbulent cascade by a Fokker–Planck equation. Physical Review Letters 78, 863.

Friedrich, R., Peinke, J., Renner, C., 2000. How to quantify deterministic and random influences on the statistics of the foreign exchange market. Physical Review Letters 84, 5224.

Friedrich, R., Peinke, J., Sahimi, M., Tabar, M.R.R., 2011. Approaching complexity by stochastic methods: from biological systems to turbulence. Physics Reports 506, 87.

Gardiner, C.W., 1997. Handbook of Stochastic Methods. Springer, Germany.

Gardner, M.W., Dorling, S.R., 1998. Artificial neural networks (the multilayer perceptron). Atmospheric Environment 32, 2627–2636.

Gardner, M., Dorling, S., 2000. Statistical surface ozone models: an improved methodology to account for non-linear behaviour. Atmospheric Environment 34, 2134.

Ghasemi, F., Sahimi, M., Peinke, J., Friedrich, R., Jafari, G.R., Tabar, M.R.R.m, 2007. Markov analysis and Kramers–Moyal expansion of nonstationary stochastic processes with application to the fluctuations in the oil price. Physical Review E 75, 060102.

Gradišek, J., Friedrich, R., Govekar, E., Grabec, I., 2003. Examples of analysis of stochastic processes based on time series data. Meccanica 38, 33–42.

Haykin, S., 1999. Neural Networks – a Comprehensive Foundation, 2 ed. MacMillan (College Publishing Company, New York.

Hooyberghs, J., Mensink, C., Dumont, G., Fierens, F., Brasseur, O., 2005. A neural network forecast for daily average PM10 concentrations in Belgium. Atmospheric Environment 39, 32793289.

Kleinhans, D., Friedrich, R., Nawroth, A., Peinke, J., 2005. An iterative procedure for the estimation of drift and diffusion coefficients of Langevin processes. Physical Letters A 346, 42–46.

Kolehmainen, M., Martikainen, H., Ruuskanen, J., 2001. Neural networks and periodic components used in air quality forecasting. Atmospheric Environment 35, 815–825.

Kukkonen, J., Partanen, L., Karppinen, 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 NO_2 and PM10 concentrations, compared with a deterministic modeling system and measurements in central Helsinki. Atmospheric Environment 37, 45394550.

Lade, S.J., 2009. Finite sampling interval effects in KramersMoyal analysis. Physical Letters A 373, 3705–3709.

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.

Lind, P.G., Mora, A., Gallas, J.A.C., Haase, M., 2005. Reducing stochasticity in the North Atlantic oscillation with coupled Langevin equations. Physical Review E 72, 056706.

Lind, P.G., Mora, A., Haase, M., Gallas, J.A.C., 2007. Minimizing stochasticity in the NAO index. International Journal of Bifurcation and Chaos 17 (10), 3461–3466.

Lind, P.G., Haase, M., Boettcher, F., Peinke, J., Kleinhans, D., Friedrich, R., 2010. Extracting strong measurement noise from stochastic time series: applications to empirical data. Physical Review E 81, 041125.

Luecken, D.J., Hutzell, W.T., Gipson, G.L., 2006. Development and analysis of air quality modeling simulations for hazardous air pollutants. Atmospheric Environment 40, 5087–5096.

Middleton, D., 1997. A new model to forecast urban air quality: BOXURB. Environmental Monitoring and Assessment 52, 315335.

Nejadkoorki, F., Baroutian, S., 2012. Forecasting extreme PM_{10} concentrations using artificial neural networks. International Journal of Environmental Research 6, 277–284.

Papanastasiou, D.K., Melas, D., Kioutsioukis, I., 2007. Development and assessment of neural network and multiple regression models in order to predict PM10 levels in a medium-sized Mediterranean city. Water, Air and Soil Pollution 182, 325–334.

Pearson, K., 1901. On lines and planes of closest fit to systems of points in space. Philosophical Magazine 2, 559.

Perez, P., 2001. Prediction of sulfur dioxide concentrations at a site near downtown Santiago, Chile. Atmospheric Environment 35, 49294935.

Perez, P., Trier, A., Reyes, J., 2000. Prediction of PM2.5 concentrations several hours in advance using neural networks in Santiago, Chile. Atmospheric Environment 34, 11891196.

Press, W.H., Flannery, B.P., Teukolsky, S.A., Vetterling, W.T., 1992. Numerical Recipes. Cambridge University Press, Cambridge.

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

Raischel, F., Scholz, T., Lopes, V.V., Lind, P.G., 2013. Uncovering Wind Turbine Properties through Two-dimensional Stochastic Modeling of Wind Dynamics. Submitted 2013.

Reich, S., Gomez, D., Dawidowski, L., 1999. Artificial neural network for the identification of unknown air pollution sources. Atmospheric Environment 33, 30453052.

Risken, H., 1984. The Fokker–Planck Equation. Springer, Heidelberg.

Shi, J.P., Harrison, R.M., 1999. Regression modelling of hourly NO_x and NO_2 concentrations in urban air in London. Atmospheric Environment 31 (24), 40814094.
Sokhi, R.S., San José, R., Kitwiroon, N., Fragkou, E., Pérez, J.L., Middleton, D.R., 2006. Prediction of ozone levels in London using the MM5CMAQ modelling system. Environmental Modelling & Software 21, 566576.
Trigo, R., DaCamara, C., 2000. Circulation weather types and their impact on the precipitation regime in Portugal. International Journal of Climatology 20, 15591581.
Trigo, R., Palutikof, J., 1999. Simulation of daily temperatures for climate change scenarios over Portugal: a neural network model approach. Climate Research 13, 4559.
van Mourik, A.M., Daffertshofer, A., Beek, P.J., 2006. Deterministic and stochastic features of rhythmic human movement. Biological Cybernetics 94, 233–244.
Vasconcelos, V., Raischel, F., Haase, M., Peinke, J., Wächter, M., Lind, P.G., Kleinhans, D., 2011. Principal axes of stochastic motion. Physical Review E 84, 031103.
Weisberg, S., 1985. Applied Linear Regression. John Wiley & Sons, New York.
Wilks, D., 2006. Statistical Methods in the Atmospheric Sciences. No. 59 in International Geophysics, second ed. Academic Press.
Yi, J., Prybutok, 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.