Physics Letters A 376 (2012) 2081–2089

Contents lists available at SciVerse ScienceDirect

Physics Letters A

www.elsevier.com/locate/pla

Searching for optimal variables in real multivariate stochastic data

F. Raischel [a,*], A. Russo [b], M. Haase [c], D. Kleinhans [d,e], P.G. Lind [a,f]

[a] Center for Theoretical and Computational Physics, University of Lisbon, Av. Prof. Gama Pinto 2, 1649-003 Lisbon, Portugal
[b] Center for Geophysics, IDL, University of Lisbon, 1749-016 Lisbon, Portugal
[c] Institute for High Performance Computing, University of Stuttgart, Nobelstr. 19, D-70569 Stuttgart, Germany
[d] Institute for Biological and Environmental Sciences, University of Gothenburg, Box 461, SE-405 30 Göteborg, Sweden
[e] Institute of Theoretical Physics, University of Münster, D-48149 Münster, Germany
[f] Departamento de Física, Faculdade de Ciências da Universidade de Lisboa, 1649-003 Lisboa, Portugal

ARTICLE INFO

Article history:
Received 9 November 2011
Received in revised form 7 May 2012
Accepted 9 May 2012
Available online 11 May 2012
Communicated by C.R. Doering

Keywords:
Stochastic systems
Environmental research
Pollutants
Langevin equation

ABSTRACT

By implementing a recent technique for the determination of stochastic eigendirections of two coupled stochastic variables, we investigate the evolution of fluctuations of NO_2 concentrations at two monitoring stations in the city of Lisbon, Portugal. We analyze the stochastic part of the measurements recorded at the monitoring stations by means of a method where the two concentrations are considered as stochastic variables evolving according to a system of coupled stochastic differential equations. Analysis of their structure allows for transforming the set of measured variables to a set of derived variables, one of them with reduced stochasticity. For the specific case of NO_2 concentration measures, the set of derived variables are well approximated by a global rotation of the original set of measured variables. We conclude that the stochastic sources at each station are independent from each other and typically have amplitudes of the order of the deterministic contributions. Such findings show significant limitations when predicting such quantities. Still, we briefly discuss how predictive power can be increased in general in the light of our methods.

1. Introduction

The industrial and urban development during the last decades has led to a general decrease of air quality, drastically affecting urban environmental and human life quality. Although, according to the European Environment Agency report [1], air quality has improved in general during the last years, this enhancement was not significant enough to ensure good air quality in all urban areas. One of the pollutants with negative impact on health and environment is NO_2. Anthropogenic NO_2 is mainly emitted by vehicles and industrial processes. NO_2 has not only severe effects on health causing e.g. respiratory and cardiovascular diseases, it also affects the environment [2] as nitrogen deposition leads to eutrophication [3]. A better understanding of the mechanisms that influence production, transport, and decomposition of NO_2 is therefore important. Previous studies revealed that temperature, wind speed and direction, relative humidity, cloud cover, dew point temperature, sea level pressure, precipitation, and mixing layer height are relevant meteorological variables to model the concentrations of air pollutants [2,4–7]. In particular, approaches that deal with the evolution of the NO_2 concentration at individual city locations are important for forecasting the air quality of urban regions.

Recently a framework [9,10] for analyzing measurements on complex systems was introduced, aiming for a quantitative estimation of drift and diffusion functions from measured data. These functions can be identified with the deterministic and stochastic contributions to the dynamics, respectively, and give a considerable insight into the underlying systems. The framework was already successfully applied for instance to describe turbulent flows [9] and the evolution of climate indices [11,28], stock market indices [12], and oil prices [13]. At the same time, the basic method has been refined in particular with respect to the impact of finite sampling effects [14,15], the impact of measurement noise [16–18], and the role of local eigendirections of the diffusion matrices [19].

In this Letter, we aim to apply some recent methods for deriving variables with reduced stochastic fluctuations to empirical data. Namely, we adapt this framework for analyzing measurements of NO_2 concentrations in the metropolitan region of Lisbon, Portugal (see Fig. 1), taken over several years. We argue that the temporal fluctuations of these concentrations result from two independent contributions: one periodic and one stochastic. The periodic part describes daily, weekly, seasonal and yearly variations of the concentration, which is an accepted and well-studied result [8].

* Corresponding author.
E-mail address: raischel@cii.fc.ul.pt (F. Raischel).

0375-9601/$ – see front matter
http://dx.doi.org/10.1016/j.physleta.2012.05.017

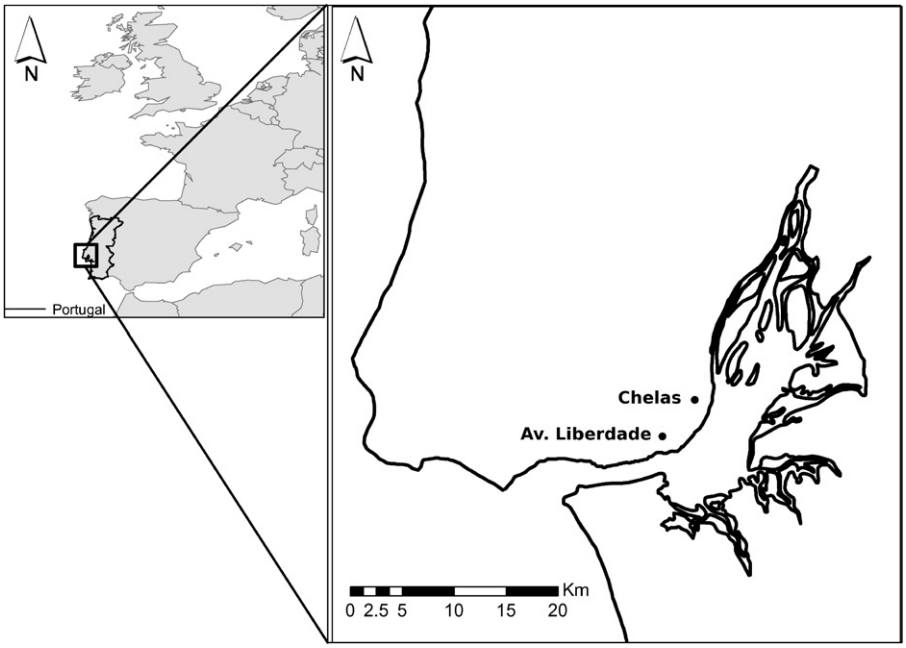

Fig. 1. NO_2 measurement stations in the region of Lisbon (Portugal) at the Southwestern coast of Europe. In this Letter we focus on the set of measurements taken at the stations of Chelas and Avenida da Liberdade with approximately 10^5 data points each extracted in the period between 1995 and 2006.

The stochastic contribution can be modelled through a stochastic differential equation [22] having two terms, one drift forcing (deterministic) and one diffusive fluctuation (stochastic).

When addressing stochastic higher-dimensional systems it is typically difficult to identify the variables most relevant for a proper description of the system's evolution. In geophysical applications, the reduction of the full set of variables to only a few variables often is achieved by means of the so-called Principal Component Analysis (PCA) [20] or other standard reduction methods, such as stepwise regression or ARIMA [21]. However, the inherent fluctuations are not so commonly investigated.

In this Letter we will apply a recent method for reconstructing the phase space of two stochastic variables, which evolve according to a set of two coupled stochastic equations defined through drift vectors and diffusion matrices [19]. The method is based on the eigenvalues of the diffusion matrices, from which it is possible to derive a path in phase space through which the deterministic contribution is enhanced. This technique implies a transformation of variables and allows for the investigation of the minimal number of independent sources of stochastic forcing in the system. In particular, a rather small eigenvalue of the diffusion matrix, compared to the average value of all the other, corresponds to one eigendirection in which stochastic fluctuations may be neglected, reducing the number of stochastic variables taken for describing the system's evolution. On the contrary, having all eigenvalues of the same order of magnitude means that the number of independent stochastic forces equals the number of variables. Moreover, as we will see, a direct inspection of the diffusion functions enables one to ascertain if the stochastic contributions, one for each variable, are coupled among them or not. Therefore, we argue that the diagonalization of the diffusion matrices gives insight into the system.

We start in Section 2 by describing the properties and the preparation of the data set. Consecutively, in Sections 3 and 4, the modeling of the time series as a Langevin process is carried out and its transformation to a new coordinate system are described in Sections 5 and 6, respectively. In Section 7 we discuss the performance of the transformation of the coordinates obtained by our approach compared to other techniques commonly used for statistical analysis of measured data. Section 8 closes this Letter with a general summary and ideas on the interpretation of the trans-

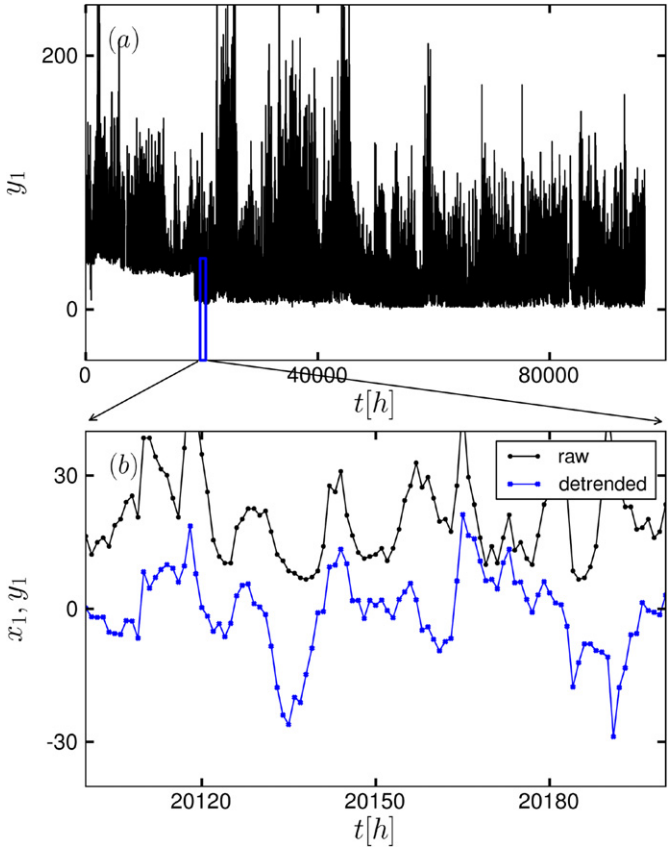

Fig. 2. (a) Time series of the NO_2 concentration at the station of Chelas, before detrending according to Eq. (2), and (b) a zoom-in of these "raw" time series y_1 compared to the detrended series x_1, which takes averages of 52-weeks periods, and then a second detrend with daily averages. Vertical offset of same plots are done for clarity. For the station at Avenida da Liberdade similar features are found (not shown).

formed time series with respect to the underlying environmental processes.

2. NO_2 measurements in Lisbon

In this section we briefly describe the sets of data analyzed in this Letter as well as its preparation for analyzing the stochastic components of the measurements.

The data set covers hourly measurements of NO_2 concentration, taken at 22 stations in the urban center of Lisbon recorded from of 1995 to 2006. For this study we choose the data from 1995 to 2005 for the monitoring stations at Chelas and at Avenida da Liberdade. These stations are located at a distance of ~ 4 km from each other, see Fig. 1. In the following, the NO_2 concentrations at the stations of Chelas and Avenida da Liberdade will be designated as $y_1(t)$ and $y_2(t)$, respectively, omitting the temporal dependency when not necessary.

Increments in time are always of 1 hour. Each of the data sets contains 10^5 measurement points approximately, including some periods of incomplete or erroneous measurements that are disregarded for our analysis. In the case of the chosen stations, the series of measurements y_1 and y_2 contain 3726 and 4548 instances of measurement errors, respectively.

The concentration of NO_2 is strongly driven by daily, weekly, monthly and yearly anthropogenic routines, and also by periodic atmospheric processes. For instance the rush hours on working days have an almost immediate impact on the NO_2 concentration, and thus, on air quality. The 24 hours and one week cycles are both traffic related and mirror daily and weekly cycles. The measurements of NO_2 are therefore influenced by different periodic forcings and, since we are interested in the fluctuations of NO_2 concentrations, the periodic behavior must be first detrended. The detrended series for y_1 and y_2, represented below as x_1 and x_2, respectively, are obtained as follows.

One first partitions the data in segments of length N, which we suppose to be a multiple of relevant periodic fluctuations in the data set. As a second step, a mean segment is calculated by averaging measurements with the same position in the segment over the entire data set according to

$$\mathcal{N}_i(n) \equiv \left\langle y_i(t) \,\middle|\, t = n + mN,\ m = 0, 1, \ldots \right\rangle \tag{1}$$

for $n = 0, 1, \ldots, N-1$. The detrended data set x_i is then calculated by subtracting the respective values of the mean segment from the measured data,

$$x_i(t) \equiv y_i(t) - \mathcal{N}_i(t \bmod N) \tag{2}$$

for $t = 1, \ldots, T$ with $T > N$. If T is the size of the data set, our simulations have shown that averages over $N = 52$ weeks is the best choice for the entire data set, to take into account all known periodicities mentioned above. With this detrending method, some periodicities with variable phase remain. To filter also these periodicities, a second detrending with $N = 1$ day is then performed on consecutive periods of $T = 14$ days.

Fig. 2a shows the original data y_1 for the station of Chelas. A zoom-in of a small time interval is plot in Fig. 2b together with the corresponding detrended data x_1. From now on, if not stated explicitly otherwise, we will only consider the detrended time series x_1 and x_2. Next describe their characteristics by means of a stochastic process.

3. Modeling stochasticity in series of NO_2 concentrations: Langevin processes

The detrended series x_i in Eq. (2) reflect the remaining stochastic components of the measurements at the respective stations of Chelas and Avenida da Liberdade. In this section we assume that,

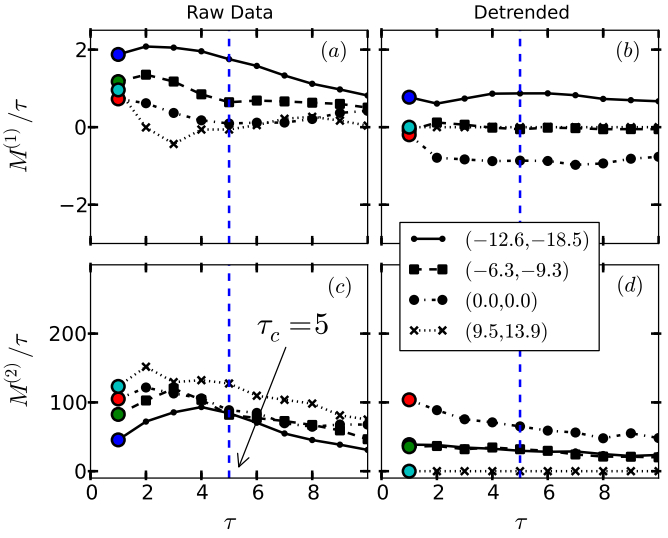

Fig. 3. First conditional moments $M^{(1)}$ for (a) the original series and (b) the detrended series, with different NO_2 concentrations (x_1, x_2) at each one of the two stations (see legend). The corresponding second conditional moments $M^{(2)}$ are shown in (c) and (d), respectively. These moments are computed according to Eqs. (9a) and (9b). While the original data presents oscillations beyond a given time interval $\tau_c \sim 5$, the detrended time series does not (see text). The value of the corresponding Kramers–Moyal coefficient at the value of (x_1, x_2) chosen is given by Eq. (8) for the lowest value of τ, i.e. one.

with two variables, the stochastic process is modeled by a system of two coupled Langevin equations, containing a deterministic and a stochastic part, described through a drift vector and a diffusion matrix, respectively.

For the general case of a K-dimensional state vector $\mathbf{X} = (x_1, \ldots, x_K)$, the Itô–Langevin equations describing the evolution of a particular trajectory in time read [22,23]:

$$\frac{d\mathbf{X}}{dt} = \mathbf{h}(\mathbf{X}) + \mathbf{g}(\mathbf{X})\boldsymbol{\Gamma}(t), \tag{3}$$

where $\boldsymbol{\Gamma} = (\Gamma_1, \ldots, \Gamma_K)$ is a set of K independent stochastic forces with Gaussian distribution fulfilling

$$\left\langle \Gamma_i(t) \right\rangle = 0, \tag{4a}$$

$$\left\langle \Gamma_i(t)\Gamma_j(t') \right\rangle = 2\delta_{ij}\delta(t - t'). \tag{4b}$$

The two terms on the right-hand side of Eq. (3) include both the deterministic contribution, $\mathbf{h} = \{h_i\}$, and the stochastic contribution, $\mathbf{g} = \{g_{ij}\}$. The deterministic contribution describes the physical forces which drive the system, while functions $\mathbf{g}$ account for the amplitudes of the different sources of fluctuations $\boldsymbol{\Gamma}$ [10].

The coefficients $\mathbf{h}$ and $\mathbf{g}$ are directly related to the drift vectors and diffusion matrices [22]

$$D_i^{(1)}(\mathbf{X}) = h_i(\mathbf{X}), \tag{5}$$

$$D_{ij}^{(2)}(\mathbf{X}) = \sum_{k=1}^{K} g_{ik}(\mathbf{X})g_{jk}(\mathbf{X}) \tag{6}$$

for $i, j = 1, \ldots, K$, describing the evolution of the joint probability density function (PDF) $f(\mathbf{X}, t)$ by means of the Fokker–Planck equation [22,23]:

$$\frac{\partial}{\partial t} f(\mathbf{X}, t) = -\sum_{k=1}^{K} \frac{\partial}{\partial x_k} D_k^{(1)}(\mathbf{X}) f(\mathbf{X}, t) + \sum_{k=1}^{K}\sum_{m=1}^{K} \frac{\partial^2}{\partial x_k \partial x_m} D_{km}^{(2)}(\mathbf{X}) f(\mathbf{X}, t). \tag{7}$$

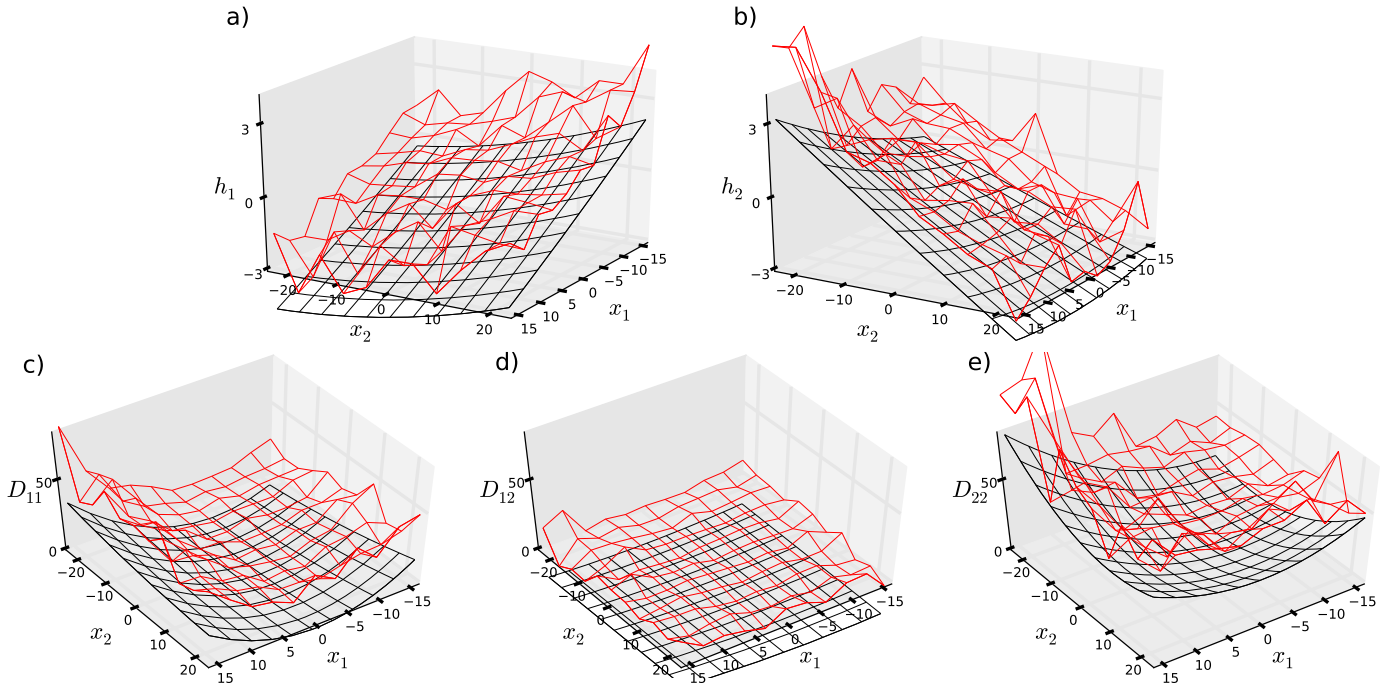

Fig. 4. For the detrended series we plot (a–b) both components of the drift vector $\mathbf{h} = (h_1, h_2)$ and (c–e) the components of diffusion matrix $\mathbf{D}^{(2)} = \{D_{ij}^{(2)}\}$. The corresponding fitted surfaces (black) are vertically offset for clarity. Since $\mathbf{D}^{(2)}$ is symmetric (see text) one has $D_{12}^{(2)} = D_{21}^{(2)}$.

As done previously in other contexts [10–14,16,17], the drift vector and the diffusion matrix can be derived directly from the data.

Statistically, the drift and diffusion coefficients $D_i^{(1)}$ and $D_{ij}^{(2)}$ are defined as

$$\mathbf{D}^{(k)}(\mathbf{X}) = \lim_{\tau \to 0} \frac{1}{\tau} \frac{\mathbf{M}^{(k)}(\mathbf{X}, \tau)}{k!}, \tag{8}$$

with the first and second conditional moments given by

$$M_i^{(1)}(\mathbf{X}, \tau) = \big\langle Y_i(t+\tau) - Y_i(t) \big| \mathbf{Y}(t) = \mathbf{X} \big\rangle, \tag{9a}$$

$$M_{ij}^{(2)}(\mathbf{X}, \tau) = \big\langle \big(Y_i(t+\tau) - Y_i(t)\big)\big(Y_j(t+\tau) - Y_j(t)\big) \big| \mathbf{Y}(t) = \mathbf{X} \big\rangle. \tag{9b}$$

Here $\langle \cdot | \mathbf{Y}(t) = \mathbf{X} \rangle$ symbolizes conditional averaging over all events that fulfill the condition $\mathbf{Y}(t) = \mathbf{X}$.

To determine the underlying Langevin equations, defined in Eq. (3), one additionally needs to solve Eqs. (5) and (6). In particular, the calculation of matrices $\mathbf{g}$ from the diffusion matrices requires to solve $\mathbf{D}^{(2)} = \mathbf{g}\mathbf{g}^T$, e.g. by means of diagonalization, $\mathbf{D}_{\text{diag}}^{(2)} = \mathbf{P}\mathbf{D}^{(2)}\mathbf{P}^{-1}$, with $\mathbf{P}$ the orthogonal matrix of eigenvectors of $\mathbf{D}^{(2)}$. The family of solutions is given by $\mathbf{g} = \mathbf{P}^T \sqrt{\mathbf{D}_{\text{diag}}^{(2)}} \mathbf{P}\mathbf{O}$, where $\mathbf{O}$ is an arbitrary orthogonal matrix, obeying $\mathbf{O}\mathbf{O}^T = 1$. The matrices $\mathbf{D}^{(2)}$ are symmetric and positive semi-definite with all their eigenvalues real and non-negative (see Eq. (9b)), and therefore $\sqrt{\mathbf{D}_{\text{diag}}^{(2)}}$ is well-defined. For any choice of $\mathbf{O}$ the analysis below does not change, and therefore we choose for simplicity $\mathbf{O}$ as the identity matrix.

The computation of the conditional moments is based on their statistical τ-dependence for small τ [10,17]. Previous works showed that Eqs. (9a) and (9b) are an operational definition of the conditional moments that can easily be implemented for the direct estimation of the drift and diffusion coefficients from the data [10,17]. In some practical situations, the limit in Eq. (8) can be approximated by the slope of a linear fit of the corresponding conditional moments at small τ. When such linear fit is not possible, an alternative estimate is to consider the first value of $M(\tau)/\tau$ at the lowest value of τ [14]. We will use this latter estimate for deriving the drift and diffusion coefficients, underlying the evolution of NO_2 concentration in Lisbon.

Within this framework, we consider the two-dimensional system of NO_2 concentrations $\mathbf{X} = (x_1, x_2)$ describing the fluctuations at the stations of Chelas and Avenida da Liberdade, see Fig. 1. In order to comply with a Langevin process, as defined in Eq. (3), we first verify that both data sets exhibit Markovian properties, which we show next for component x_1 only, for sake of clarity. For x_2 the results are similar.

As Fig. 3 indicates, the conditional moments of the time series show no evidence of measurement noise as τ approaches zero [16]: $M^{(1)}/\tau$ do not diverge when $\tau \to 0$. This is true, both before and after detrending. For τ smaller than a limiting value τ_c, some oscillations are observed in the case without detrending, although they have no impact on the estimate of the corresponding Kramers–Moyal coefficients, as compared to our method of using the value at $\tau = 1$ as estimate. For details see Ref. [14].

The resulting components of the drift and diffusion coefficients are plotted in Fig. 4. As one sees all surfaces are adequately fitted by a quadratic polynomial

$$p(x_1, x_2) = a_1 x_1^2 + a_2 x_2^2 + a_3 x_1 x_2 + a_4 x_1 + a_5 x_2 + a_6, \tag{10}$$

where p denotes the drift and diffusion components, $D_i^{(1)}$ and $D_{ij}^{(2)}$, respectively, and the coefficients a_i are computed from a least-square procedure on the drift and diffusion components as functions of the detrended variables x_1 and x_2.

4. Analysis of Markov properties for the series of NO_2 concentrations

The Markovian nature of the variable x_1 can be investigated by considering the differences between the conditional one-point probability $p(x_1^{(1)}, t | x_1^{(2)}, t - \tau_2)$ and the conditional two-point

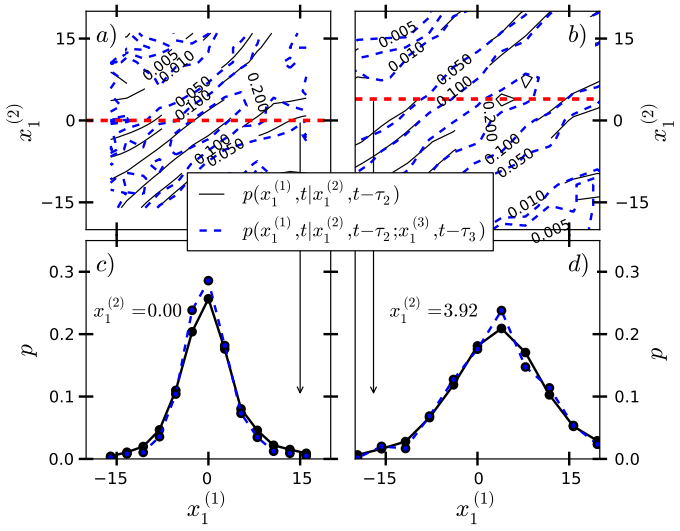

Fig. 5. Contour plots of conditional probabilities (solid curves) and conditional two-point probabilities (dashed curves) computed from the detrended time series x_1 and x_2 with $\tau_2 = 1$ hour for (a) $\tau_3 = 2$ hours and (b) $\tau_3 = 10$ hours. The corresponding cuts through contour planes, indicated by the horizontal dashed lines, are shown in (c) and (d) with a good matching between the respective one-point and two-point conditional probabilities. The distributions were computed with 13 bins for each variable using a sample of 10^5 data points.

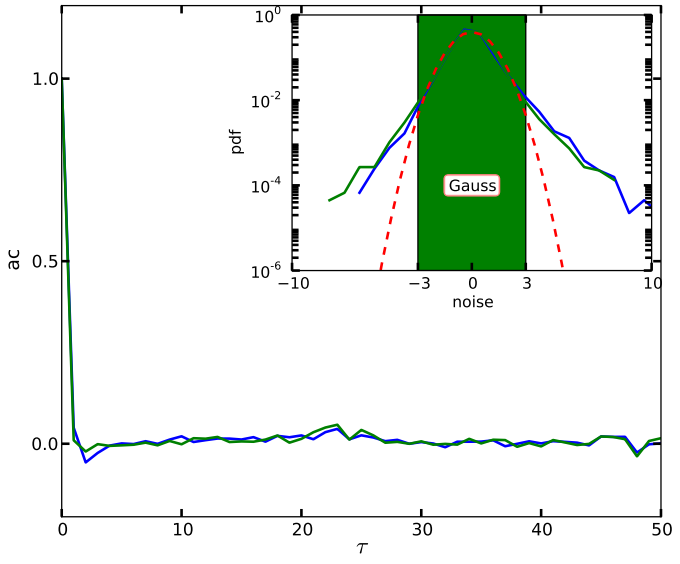

Fig. 6. Autocorrelation of the reconstructed dynamical noises Γ_1, Γ_2 (stochastic fluctuation), indicating that they are δ-correlated. The inset shows the probability density function (PDF) of the reconstructed noise normalized to variance 1 (lines) and a normal distribution for comparison (dashed line).

probability $p(x_1^{(1)}, t|x_1^{(2)}, t-\tau_2; x_1^{(3)}, t-\tau_3)$. If the process is Markovian on time scales larger than τ_2, then these probability distributions should not differ significantly [10] for any choice of τ_3. Indeed, as can be seen from Fig. 5, the Markovian properties seem to be fulfilled for $\tau_2 = 1$ h and both $\tau_3 = 2$ h and $\tau_3 = 10$ h. We therefore observe strong indications that the process is Markovian already at the sampling rate of the data points of 1 h and for time lags longer than 1 h.

Further, it is also necessary to check the Gaussian nature of the stochastic force Γ and ascertain it indeed obeys Eqs. (4). Using the measured time series and the estimated KM-coefficients, the noise $\boldsymbol{\Gamma}(t)$ can be reconstructed from a numerical discretization of Eq. (3) solved with respect to Γ [24], namely solving

$$\boldsymbol{\Gamma} = \mathbf{g}^{-1}(\mathbf{X})\big(\bar{\mathbf{X}} - \mathbf{h}(\mathbf{X})\big), \tag{11}$$

where $\bar{\mathbf{X}} = \mathbf{X}(t+1) - \mathbf{X}(t)$ and $\mathbf{h}(\mathbf{X})$ and $\mathbf{g}(\mathbf{X})$ are evaluated at $\mathbf{X} = \mathbf{X}(t)$.

The resulting noise is analyzed with respect to its autocorrelation, shown in Fig. 6: the autocorrelation decays to zero for the very first values of τ, which strongly supports to treat Γ_1 and Γ_2 as a white, δ-correlated noise source.

For ascertaining the Gaussian nature of the stochastic sources we plot in the inset of Fig. 6 the PDF of the reconstructed noise time series Γ_1 and Γ_2 (solid lines) against a Gaussian distribution (dashed lines).

As one sees from the inset, in the range comprising over 95% of the Gaussian noise, the distributions for the stochastic sources are well approximated by a Gaussian distribution. We find it reasonable to assume, therefore, that the data series can be approximated sufficiently well by a Fokker–Planck equation. The deviations observed for the extreme values, are common in the analysis of long-term field measurements, showing tails for close to exponential decay.

From the tests described in this section one may satisfactorily take the series x_1 and x_2 as a set described by two coupled Langevin equations, Eq. (3) with $K = 2$. Next we derive these equations from the sets of measurements x_1 and x_2.

5. Deriving optimal variables: eigensystem for NO_2 measurements at different stations

Having successfully determined the drift and diffusion constants describing the respective deterministic forcing and stochastic fluctuations of the system of NO_2 concentration measurements, we now determine the eigensystem of the diffusion matrices and investigate its principal directions. This procedure was described in detail in [19] and was previously applied to a two-dimensional sub-critical bifurcation [25] and to the analysis of human movement [26]. It will be briefly outlined here, for K variables.

Diffusion matrices are numerically estimated on a mesh of points in phase space, as shown for example in Fig. 4c-e. Then at each mesh point the K eigenvalues and corresponding eigenvectors of the estimated matrices are calculated. The diffusion matrices contain information about the stochastic fluctuations acting on the system and we use the local eigensystems of the matrix for a further characterization of these forces. In particular, a vanishing eigenvalue indicates that the corresponding stochastic force may be neglected.

We are looking for a transform of the original coordinates $\mathbf{X} = \{x_i\}$ into new ones $\tilde{\mathbf{X}} = \{\tilde{x}_i\}$, such that the new coordinates are aligned in the directions of the eigenvectors of the diffusion matrix in each mesh point, i.e. the principal direction in which the diffusion matrix is diagonal. Diagonalizing the diffusion matrix decouples the stochastic contribution in the set of variables, and if the eigenvalues in the transformed coordinates are significantly different, we are able to restrict our investigation to the coordinates with lower stochasticity.

We therefore look for a two-times continuously differentiable function $\mathbf{F}$ with

$$\tilde{\mathbf{X}} = \mathbf{F}(\mathbf{X}, t), \tag{12}$$

for which [22,23] the deterministic and stochastic parts in the Langevin systems of equations, transform respectively as [19]

$$\tilde{h}_i^{(1)}(\tilde{\mathbf{X}}) = \sum_{k=1}^{N}\left(h_k^{(1)}(\mathbf{X})\frac{\partial F_i}{\partial x_k} + \sum_{l=1}^{N}\sum_{j=1}^{N} g_{lj}(\mathbf{X})g_{kj}(\mathbf{X})\frac{\partial^2 F_i}{\partial x_k \partial x_l}\right), \tag{13}$$

$$\tilde{g}_{ij}(\tilde{\mathbf{X}}) = \sum_{k=1}^{N} g_{kj}(\mathbf{X})\frac{\partial F_i}{\partial x_k} \tag{14}$$

where the second equation reads $\tilde{\mathbf{g}}(\tilde{\mathbf{X}}) = \mathbf{J}(\mathbf{X})\mathbf{g}(\mathbf{X})$, with $\mathbf{J}(\mathbf{X})$ the Jacobian of our transformation $\mathbf{F}$. For reasons of clarity in the following we do not explicitly notate the dependence on $\mathbf{X}$ and $\tilde{\mathbf{X}}$.

The eigenvectors $\mathbf{u}_k$ of matrices $\tilde{\mathbf{g}}$ with coordinates in local bases $\tilde{\mathbf{e}}_i$, can be incorporated in matrices $\mathbf{U} = [\mathbf{u}_1\ \mathbf{u}_2\ \ldots\ \mathbf{u}_K]$. Defining $\tilde{\mathbf{U}}$ as $\mathbf{U} = \mathbf{J}^T\tilde{\mathbf{U}}$ one then obtains (see Eq. (14))

$$\tilde{\mathbf{U}}^T\tilde{\mathbf{g}}\tilde{\mathbf{g}}^T\tilde{\mathbf{U}} = \mathbf{U}^T\mathbf{g}^T\mathbf{g}\mathbf{U}. \tag{15}$$

By definition the inverse transform $\mathbf{F}^{-1}(\tilde{\mathbf{X}})$ is chosen such that the normalized eigenvectors are given by

$$\mathbf{u}_k = \frac{1}{s_k}\frac{\partial \mathbf{F}^{-1}}{\partial \tilde{x}_k} \tag{16}$$

with

$$s_k = \left|\frac{\partial \mathbf{F}^{-1}}{\partial \tilde{x}_k}\right|, \tag{17}$$

i.e. the respective square sum of the columns in the Jacobian of the inverse transform. Taking into account this scaling factor the eigenvalues in the new coordinate system can be calculated [19],

$$\tilde{\mathbf{D}}^{(2)} = \operatorname{diag}\left[\ \frac{\lambda_1}{s_1^2}\quad \frac{\lambda_2}{s_2^2}\quad \cdots\quad \frac{\lambda_K}{s_K^2}\ \right], \tag{18}$$

where λ_i $(i = 1, \ldots, K)$ are the eigenvalues of the diagonalized matrix $\mathbf{D}^{(2)}_{\text{diag}}$.

In general, the eigenvalues of the diffusion matrix indicate the amplitude of the stochastic force and the corresponding eigenvector indicates the direction towards which such force acts. These directions can be regarded as principal axes of the underlying stochastic dynamics [19].

In particular the vector field yielding at each mesh point the eigenvector associated to the smallest eigenvalue of matrix $\mathbf{g}$ defines the paths in phase space towards which the fluctuations are minimal. If this eigenvalue is very small compared to all the other, the corresponding stochastic force can be neglected and the system can be assumed to have only $K - 1$ independent stochastic forces, reducing the number of stochastic variables in the system.

Notice however that, whereas in a Cartesian coordinate system the eigenvalues are strictly related to the amplitude of diffusion in the corresponding eigenvectors, a non-linear transformation usually changes the metric [19,27]. In the transformed system the direction of the maximal eigenvalue is not necessarily the direction with the highest diffusion. This disparity is accounted for by the factor s_i above. A much more simple case occurs when the eigenvectors are parallel (or almost) to a fixed direction, meaning that the eigendirections at each point in phase space are the same but rotated by a constant angle. Next we address such situation.

6. Transform of NO_2 concentrations to the stochastic eigendirections

In this section we apply the procedure described previously to the two series of NO_2 concentration, x_1 and x_2 in Chelas and Avenida da Liberdade, shown in Fig. 7a. The joint PDF of both concentrations x_1 is x_2 is plotted in Fig. 7b showing the region in phase space most visited by the bivariate series (x_1, x_2). A plot of the eigenvectors of $\mathbf{D}^{(2)}$ in Fig. 7c suggests that a continuous and smooth description of the corresponding sorted eigenvalues exists. Here we place at each grid point of the phase space one ellipsoid whose major and minor axes are given by the (non-normalized) eigenvectors associated to the largest and smallest eigenvalues, respectively. Since the two eigenvalues are different, the eigenvector corresponding to the lower eigenvalue describes the direction of minimum stochasticity. The eigenvectors and eigenvalues of the diffusion matrix give locally the principal directions of stochastic fluctuations (diffusion).

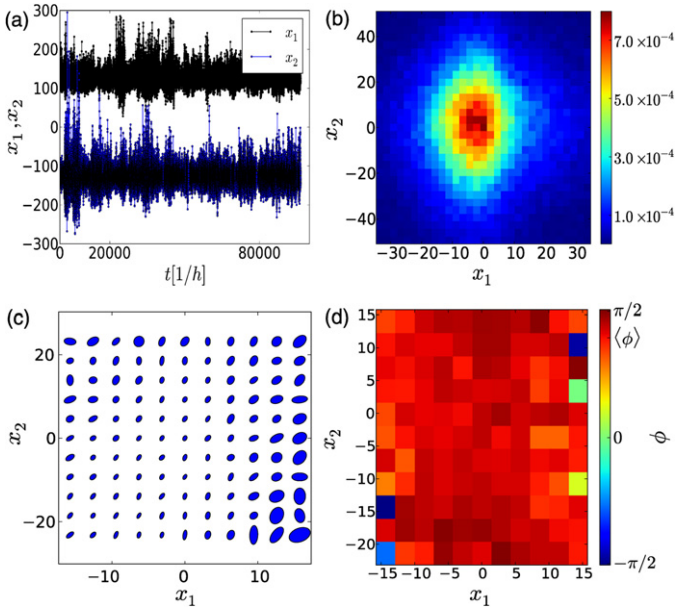

Fig. 7. (a) The original time series x_1 and x_2 and (b) their joint probability density function. In (c) we plot the two (orthogonal) eigenvectors defining the semi-axis of an ellipsoid. The major semi-axis is associated with the largest eigenvalue and correspondingly the minor semi-axis with the smallest. For each grid point in phase space (x_1, x_2) we plot (d) the angle ϕ between the eigenvector associated to the largest eigenvalue and the positive x-semi-axis. As one sees from (c) and (d) the angle ϕ does not show significant disparity in its values within the considered range, indicating a strong uncoupling between the two stations (see text). Thus, a global rotation of the axis may be considered (see Fig. 8).

In general, from a plot as the one in Fig. 7c it is possible to derive numerically the variable transformation in Eq. (12): at each grid point one determines the angle between the "largest" eigenvector and the positive horizontal axis. Fig. 7d shows the angle ϕ for the bivariate series (x_1, x_2). Rotating each ellipsoid separately by the respective ϕ-angle aligns the largest eigenvector along the horizontal direction and the smallest eigenvector along the vertical direction, yielding the two new (transformed) variables $\tilde{x}_1$ and $\tilde{x}_2$. This angle can be derived at each grid point from the corresponding diffusion $\mathbf{g}$ components namely

$$\tan 2\phi = \frac{2g_{12}}{g_{11} - g_{22}}. \tag{19}$$

The angle ϕ or its absolute value quantifies the relative off-diagonal contribution that describes the coupling of the noise terms by the diffusion matrix.

In general, what does such a transformation add to our understanding about the system? First, by definition the transformation decouples independent stochastic forces in the system. The original (detrended) pair of variables as well as the transformed pair of variable obey Eq. (3), with one important difference: the transformed pair of variables are such that each variable has a stochastic contribution governed by one independent stochastic force alone. In other words $\mathbf{g}(x_1, x_2)$ is diagonal. For the original pair of variables the stochastic contribution mixes both independent stochastic forces. Second, in a reference frame where the two independent stochastic forces are decoupled, their minimum and maximum magnitude reach the largest difference between them. In other words, one aligns the major and minor axis of the "diffusion ellipsoids" shown in Fig. 7c. In the particular case when one of the magnitudes is much smaller than other one, one of the variables can even be disregarded as a stochastic variable, reducing the number of stochastic variables describing the system. A generalization to K variables is straightforward.

Table 1
Characterizing different pairs of variables: the original pair of measures $\mathbf{y}$, the detrended pair of variables $\mathbf{x}$, the transformed pair $\tilde{\mathbf{X}}$, and, for comparison, the pair $\hat{\mathbf{x}}$ transformed according to the simpler rules in (22). For each variable or pair of variables ($i = 1, 2$) we show the mean $\langle\rangle_i$ and standard deviation σ_i of their distribution of observed values together with the rotation angle ϕ averaged over phase space, as well as the average coefficients Q and R_i for evaluating their stochastic and deterministic contributions. See Eqs. (19), (20) and (21). The correlation coefficient between both variables is also given in each case (see text).

	y	x	$\tilde{\mathbf{X}}$	$\hat{x}$
$\langle\rangle_1$	36	$5.82 \cdot 10^{-6}$	0.00935	−0.0145
$\langle\rangle_2$	64.7	$5.22 \cdot 10^{-6}$	0.0541	−0.0721
σ_1	26.8	17.3	26.7	18.3
σ_2	40.8	25.5	15.1	23.3
$\langle\|\phi\|\rangle$	1.11 ± 0.282	1.31 ± 0.203	0.117 ± 0.114	1.33 ± 0.105
$\langle Q \rangle$	0.703 ± 0.182	0.714 ± 0.103	0.679 ± 0.0946	0.593 ± 0.0445
$\langle R_1 \rangle$	0.157 ± 0.125	0.247 ± 0.167	0.19 ± 0.122	0.169 ± 0.118
$\langle R_2 \rangle$	0.183 ± 0.13	0.207 ± 0.134	0.245 ± 0.15	0.243 ± 0.153
μ	0.479	0.457	−0.254	−0.401

The value of $\langle\phi\rangle$ shown in Table 1 is the (x_1, x_2)-averaged angle $\langle\phi\rangle \simeq 0.40\pi \sim \pi/2$ indicated at the scale of Fig. 7d. Further, one inspection of Figs. 7c and 7d enables the observation that in our present case the ϕ-angle remains approximately constant at any grid point. Similar observations were made for the other pairs of stations in Lisbon (not shown). Consequently, we may conclude that for our set of stations a global rotation is enough to align the "diffusion ellipsoids". For the stations in Chelas and Avenida da Liberdade, Fig. 8a shows the result obtained after performing a global rotation by the median $\text{median}(\phi) \simeq 0.43\pi$.

In Fig. 8b both eigenvalues λ_i are plotted, corresponding to the length of the major and minor axis of the diffusion ellipsoids. While there is a significant difference between both eigenvalues, $\lambda_{\max} \sim 2.5\lambda_{\min}$ as shown in Fig. 8f, they are of the same order of magnitude. Such observation indicates the presence of two independent stochastic forces driving the bivariate signal (x_1, x_2).

The stochastic contribution for each variables of the pair (x_1, x_2) obeying Eq. (3) can be compared through one parameter Q defined at each point $\mathbf{x}$ in phase space as

$$Q^2(\mathbf{x}) = \frac{g_{11}^2(\mathbf{x}) + g_{12}^2(\mathbf{x})}{g_{21}^2(\mathbf{x}) + g_{22}^2(\mathbf{x})} \tag{20}$$

where one orders the rows of matrix $\mathbf{g}$ to guarantee $Q < 1$, i.e. variable x_1 is chosen as the one having lower stochastic contribution. When $Q = 1$ both stochastic contributions are equal. When $Q \ll 1$ one stochastic contribution can be neglected, reducing by one the number of stochastic contributions in the system. For an arbitrary number of stochastic variables, the generalization of Eq. (20) is straightforward [19].

Table 1 shows the value of coefficient Q for the set of measurements $\mathbf{y}$, for the detrended variables $\mathbf{x}$ and for the transformed detrended variables $\tilde{\mathbf{X}}$. The coefficient is averaged over the sample of points in the corresponding phase space. For $\mathbf{y}$ and $\mathbf{x}$ the smallest stochastic contribution has a magnitude of approximately 70% of the largest one, while for the transformed variables it decreases more than 2%. This magnitude is not small enough to permit neglecting one variable. We consider this finding the central result of this Letter: before transformation the pair of detrended variables include already two independent stochastic forces of the same order of magnitude.

One note is however important to stress at this point. The method applied here to empirical data deals with a transformation that operates on the diffusion matrix alone. No constraints related to the drift functions, h_1 and h_2 are considered. To evaluate the predictability of each variable i one needs to compare the total amplitude of the stochastic term with the deterministic term, namely

$$R_i^2(\mathbf{x}) = \frac{h_i^2(\mathbf{x})}{g_{i1}^2(\mathbf{x}) + g_{i2}^2(\mathbf{x})}. \tag{21}$$

Such expression is also straightforwardly extended to K variables. The larger R_i the more predictable the variable i may be, i.e. the smaller the stochastic overall contribution is compared to the deterministic part governing the evolution of the variable. In our present case, as given in Table 1, while the detrending $y \to x$ of our measurements increases the predictability of the non-periodic modes in time, the global rotation has no major effect: both coefficients R_i maintain the same order of magnitude after transform.

The correlation coefficient μ between both stations is also given in Table 1. While detrending has no significant effect on the correlation, the transform $x_i \to \tilde{x}_i$ indeed decreases its absolute value.

Figs. 8c, 8d and 8e illustrate the numerical result of each property, ϕ, Q and R for the transformed variables. Similar to such variables is the quotient between the maximum and minimum eigenvalues, shown in Fig. 8f. Similar plots are obtained for the other possible pairs of stations.

7. Comparison with standard methodologies

In this section we first address the question of how good the coordinate transform derived above is compared to other, possibly simpler transforms.

For example, we may consider a transform to coordinates which describe the mean value and difference between the two measured time series, e.g.

$$\begin{pmatrix} \hat{x}_1 \\ \hat{x}_2 \end{pmatrix} = \frac{1}{2} \begin{pmatrix} x_1 + x_2 \\ x_1 - x_2 \end{pmatrix}. \tag{22}$$

This choice is the simplest one for two variables, one describing the total amount $x_1 + x_2$ and another describing the relative amount $x_1 - x_2$. For such choice of variables we obtain a value of $Q = 0.59$, which is essentially the same as for our "optimized" variables (see Table 1), The absolute value of the angle $\langle|\phi|\rangle$ is however considerably larger than for our optimized transform, as is the correlation coefficient between the time series, meaning that this simple transform fails to decouple the noise sources. The drift-diffusion quotients yield $R_1 = 0.17$ and $R_2 = 0.24$, showing again no better predictability in comparison with the original variables.

In our case we saw that the eigendirections do not depend much on the detrended variables x_1 and x_2, which implies that they are functionally decoupled. However, sometimes it is necessary to consider a proper scaling of the variables [19]. In such cases, we find it advisable to use a more general transform to generalized polar coordinates given by

$$\begin{pmatrix} \tilde{x}_1 \\ \tilde{x}_2 \end{pmatrix} = \begin{pmatrix} r_g \\ \theta_g \end{pmatrix} = \begin{pmatrix} \sqrt{(\alpha x_1 + \beta)^2 + (\gamma x_2 + \delta)^2} \\ \arctan(\frac{\gamma x_2 + \delta}{\alpha x_1 + \beta}) + \epsilon \end{pmatrix} \tag{23}$$

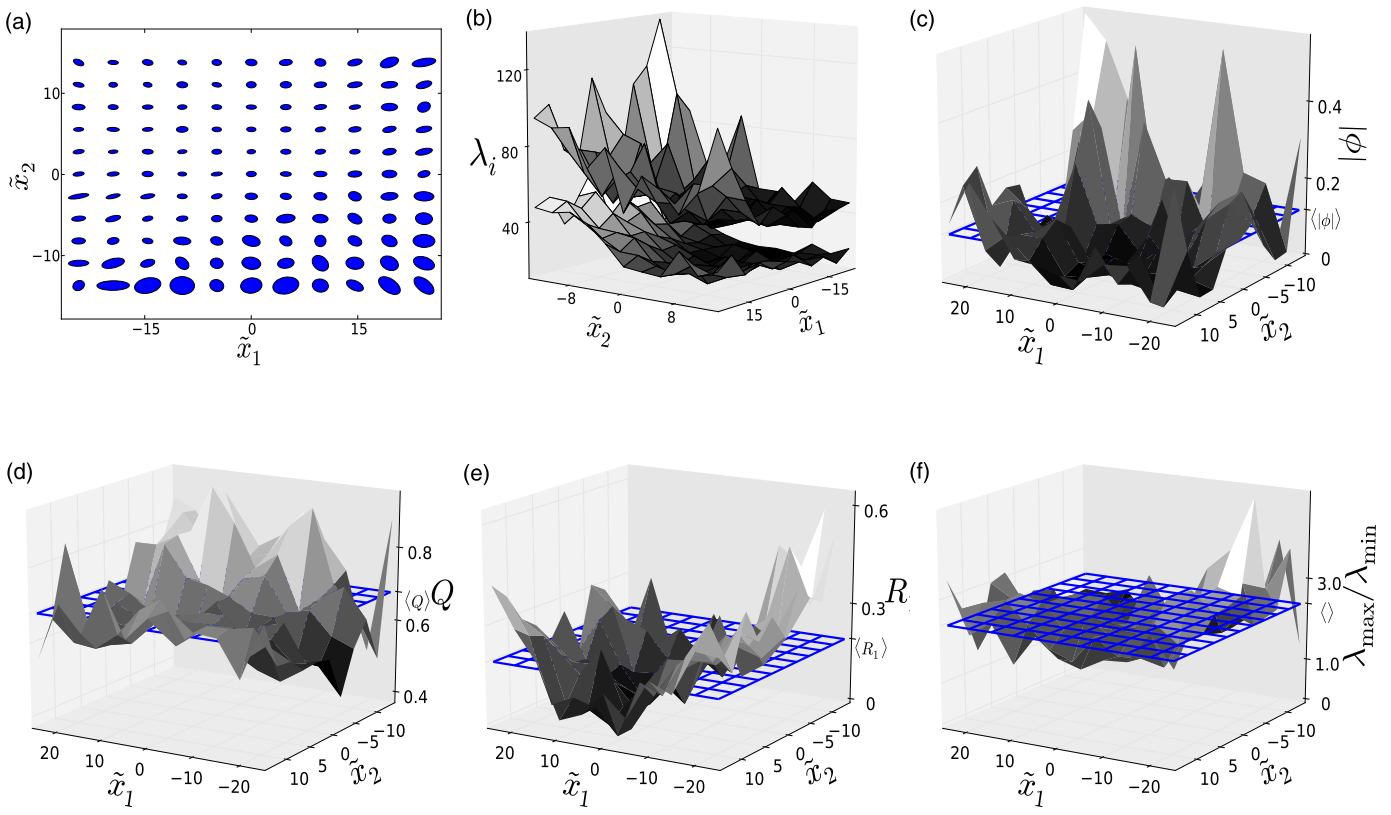

Fig. 8. Transformed variables through a global rotation of the x_1 and x_2 axis (check Fig. 7) into new variables, $\tilde{x}_1$ and $\tilde{x}_2$. (a) Eigenvectors of the transformed variables and (b) its eigenvalues derived for the diffusion matrix of the transformed variables, together with quantities to evaluate some underlying properties of the system, namely (c) the rotation angle ϕ (see Eq. (19)), (d) the asymmetry of the stochastic influence at each variable, given by Q in Eq. (20), (e) the deterministic coefficient R in Eq. (21) and (f) the quotient between the largest λ_{max} and the smallest λ_{min} for each grid point in the transformed phase position. For each property the average value is showed with horizontal surfaces and an explicit indication at the vertical axis.

where in general the radial and angle variables, r_g and θ_g, are functions of the detrended variables x_1 and x_2. This approach has the advantage that the inverse transform $\mathbf{F}^{-1}$ is given by the simple form

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} \frac{1}{\alpha}[r_g \cos(\theta_g - \epsilon) - \beta] \\ \frac{1}{\gamma}[r_g \sin(\theta_g - \epsilon) - \delta] \end{pmatrix}. \tag{24}$$

In addition, the metric factors introduced by the polar transform can result in a more pronounced separation of the eigenvalues in the transformed coordinates.

The other question addressed within this section is the comparison between methods applied to choose the most appropriate inputs. Theoretically, any set of input data can be fed into a model for training and evaluation. However, the number of possible variables to be used and the number of ways they can be presented is too diverse to test all possible combinations. A number of statistical methods can be applied in order to choose the most appropriate set of predictors or inputs. Examples are, among other, stepwise regression, PCA, cluster analysis and ARIMA. For details, see Ref. [21] and references therein. Such pre-processing procedures, reduce the number of input variables into the models, thus eliminating the redundant information. In these standard procedures, the selection of variables is usually made independently for each monitoring station.

Another possible way to tackle redundancy is the pre-processing of data consisting of the computation of backward stepwise regressions (BSR) conducted between a target variable and all the other data sets. Based on the available common period data sets, one constructs a collection of records, composing the input vector, which includes the meteorological variables, air pollutant concentrations, etc., and together with it assumes the corresponding target, which in our case is the atmospheric concentration of a certain pollutant. Subsequently, one retains the smallest subset of statistically significant variables to predict a certain pollutant concentration automatically at a given monitoring station. In addition, BSR allows the determination of the best time lags for each input variable, typically daily and weekly cycles.

The referred techniques also allow the comparison between the original data sets and surrogate data sets including only the stochastic component. The stochastic component may be determined through a rough approximation of a mathematical function (e.g., $\sin x$), or, for example, by the presented framework. After the selection of variables and the determination of cyclic and stochastic behaviors on each time series, linear and non-linear models can be applied in order to model air pollution in each monitoring station. The forecasting capabilities of the different approaches can then be compared. Such models are also applied to each decoupled time series in order to predict next days air quality at each monitoring station. The applications of this framework, however, allows to determine the stochastic component on a efficient manner, enhancing air quality predictions.

8. Discussion and conclusion

In this Letter, we investigated the stochastic properties of a set of two simultaneous series, obtained by introducing a proper detrending of NO_2 measurements, which is able to remove periodic modes in the series. We focused in the measurements at two different stations out from a set of 22 stations in Lisbon.

Based on validity tests we assumed, that the time series after detrending were properly modeled by a system of Langevin equations. The validity of this assumption is discussed in Section 3, showing that the data sets obey the Markov property to a sufficient extent. The stochastic fluctuations show good resemblance with δ-correlated Gaussian noise.

Calculating the eigenvalues of the diffusion matrices, we found a transform that leads to a description in which the diffusion matrices are diagonal. Since the transformed variables are derived directly from the transformation that diagonalizes the diffusion matrices, they correspond to the orthogonal directions in phase space in which fluctuations are stronger (larger eigenvalue) and weaker (smaller eigenvalue), respectively.

Comparison between original and transformed variables showed that the two detrended variables are driven by stochastic forces almost decoupled from each other, showing an almost constant rotation angle of the "diffusion ellipsoid" at each point of phase space. Further, both stochastic sources have amplitudes of the order of the deterministic terms, indicating a short horizon of predictability. This procedure worked out well for the NO_2 data, since the transformation of variables resulted in decoupling the diffusion components in the new coordinates. Other transformation could be considered. For instance, we discussed how this approach could be applied for other data sets in which the diffusion ellipsoids do not align in phase space, but instead depend in non-trivial functional of the variables. In this case, the transform maps the detrended variables into two polar-like coordinates.

One question that should be addressed in a forthcoming study is to present a systematic overview on all pairs of stations studied by us in this scope but not shown thoroughly, since it was out of our main purposes. Doing that one would be able to compare in detail the results obtained through the method applied in this Letter with standard methods used for forecasting NO_2 concentration at a specific spot in the city of Lisbon.

Acknowledgements

The authors thank DAAD and FCT for financial support through the bilateral cooperation DREBM/DAAD/03/2009. F.R. (SFRH/BPD/65427/2009) and P.G.L. (*Ciência 2007*) thank Fundação para a Ciência e a Tecnologia for financial support, also with the support Ref. PEst-OE/FIS/UI0618/2011.

References

[1] EEA – European Environment Agency, The European environment. State and outlook 2010: synthesis, http://dx.doi.org/10.2800/45773, 2010.
[2] N. de Nevers, Air Pollution Control Engineering, McGraw–Hill Companies, 2000.
[3] Report "Health aspects of air pollution with particulate matter, ozone and nitrogen dioxide", World Health Organization, Denmark, 2004.
[4] M. Demuzere, R. Trigo, V. Arellano, N. van Lipzig, Atmospheric Chemistry and Physics 9 (2009) 2695.
[5] M. Gardner, S. Dorling, Atmospheric Environment 33 (1999) 709.
[6] J. Hooyberghs, C. Mensink, G. Dumont, F. Fierens, O. Brasseur, Atmospheric Environment 39 (2005) 3279.
[7] J. Kukkonen, L. Partanen, A. Karppinen, J. Ruuskanen, H. Junninen, M. Kolehmainen, H. Niska, S. Dorling, T. Chatterton, R. Foxall, G. Cawley, Atmospheric Environment 37 (2003) 4539.
[8] M. Kolehmainen, H. Martikainen, J. Ruuskanen, Atmospheric Environment 35 (2001) 815.
[9] R. Friedrich, J. Peinke, Phys. Rev. Lett. 78 (1997) 863.
[10] R. Friedrich, J. Peinke, M. Sahimi, M.R.R. Tabar, Phys. Rep. 506 (2011) 87.
[11] P.G. Lind, A. Mora, J.A.C. Gallas, M. Haase, Phys. Rev. E 72 (2005) 056706.
[12] R. Friedrich, J. Peinke, Ch. Renner, Phys. Rev. Lett. 84 (2000) 5224.
[13] F. Ghasemi, M. Sahimi, J. Peinke, R. Friedrich, G.R. Jafari, M.R.R. Tabar, Phys. Rev. E 75 (2007) 060102.
[14] D. Kleinhans, R. Friedrich, A. Nawroth, J. Peinke, Phys. Lett. A 346 (2005) 42.
[15] S.J. Lade, Phys. Lett. A 373 (2009) 3705.
[16] F. Boettcher, J. Peinke, D. Kleinhans, R. Friedrich, P.G. Lind, M. Haase, Phys. Rev. Lett. 97 (2006) 090603.
[17] P.G. Lind, M. Haase, F. Boettcher, J. Peinke, D. Kleinhans, R. Friedrich, Phys. Rev. E 81 (2010) 041125.
[18] J. Carvalho, F. Raischel, M. Haase, P.G. Lind, J. Phys. 285 (2011) 012007.
[19] V. Vasconcelos, F. Raischel, M. Haase, J. Peinke, M. Wächter, P.G. Lind, D. Kleinhans, Phys. Rev. E 84 (2011) 031103.
[20] K. Pearson, Philosophical Magazine 2 (1901) 559.
[21] D. Wilks, Statistical Methods in the Atmospheric Sciences, 2nd ed., International Geophysics, vol. 59, Academic Press, 2006.
[22] H. Risken, The Fokker–Planck Equation, Springer, Heidelberg, 1984.
[23] C.W. Gardiner, Handbook of Stochastic Methods, Springer, Germany, 1997.
[24] C. Micheletti, G. Bussi, A. Laio, J. Chem. Phys. 129 (2008) 074105.
[25] J. Gradišek, R. Friedrich, E. Govekar, I. Grabec, Meccanica 38 (2003) 33.
[26] A.M. van Mourik, A. Daffertshofer, P.J. Beek, Biological Cybernetics 94 (2006) 233.
[27] S.J. Lade, Phys. Rev. E 80 (2009) 031137.
[28] P.G. Lind, A. Mora, M. Haase, J.A.C. Gallas, Int. J. Bif. Chaos 17 (10) (2007) 3461.