Finding Trends of Airborne Harmful Pollutants by Using Recurrence Quantification Analysis

Marco A. Aceves-Fernandez.[*], Juan M. Ramos-Arreguín, J. Carlos Pedraza-Ortega, Artemio Sotomayor-Olmedo., Saúl Tovar-Arriaga

Universidad Autonoma de Queretaro, Campus Juriquilla, Av. de las Ciencias S/N, Juriquilla, Delegación Santa Rosa Jauregui, CP 76230, Querétaro, México

Abstract　In this work, the use of Recurrence Plots and Recurrence Quantification Analysis explores the changes in the non-linear behavior of harmful airborne particle concentration in four sites around London simultaneously. This research has been carried out for 6 years, using large datasets of raw data (hourly) for harmful particles such as CO, SO_2, NO_2, NO and Particulate Matter (PMx). Recurrence analysis has been shown to be a useful tool in many disciplines to find trends, rates and predictions. Nevertheless, it has not been shown before the feasibility of using these algorithms to extract information for pollution monitoring and control. Also, observations are made with the results and conclusions drawn from these observations, showing the feasibility of this approach in finding trends of airborne pollution.

Keywords　Recurrence Plot, Air Quality, Air Pollution Modelling, Atmospheric Pollution, RQA

1. Introduction

The states in nature typically change in time. The importance in the investigation of these changes in complex systems helps to understand and describe such changes. A relatively new method based on non-linear data analysis has become popular to describe the changes of these systems. This method is called recurrence plot (Eckmann, 1987; Tanio, 2009).

Recurrence-based methods have a potential for representation of measurements from complex systems. However, it is necessary to determine the time intervals and state space subsets in which the stationary assumptions are hold (Yang et al., 2011).

This contribution makes the first approach in quantify and analyze the non-linear behavior of harmful airborne particles at various sites at London, England using recurrence features embedded in the raw datasets.

2. Background

2.1. Urban Airborne Pollution

In recent times, urban air pollution has been a growing problem especially for urban communities. Size, shape and chemical properties govern the lifetime of particles in the atmosphere and the site of deposition within the respiratory tract. Health effects differ upon the size of airborne particulates (Yin et al., 2010).

Air pollution has become a real concern, particularly in large urban locations (Kilabuko et al., 2007; Mirasgedis, 2008). Also, air pollution has been held responsible for various health disorders, especially respiratory complications resulting in an increase in the number of asthmatic cases and hospital admissions in some parts of the world and has been widely documented (Liu, 2011; Weinmayr et al., 2010; Arbex et al., 2010; Guo et al., 2010).

In this contribution, five airborne particles have been chosen mainly due to their impact on human health and data availability at the proposed sites. The datasets are separated according to month of the year and type of particle. There is one data for each hour, for each particle for all four London's sites, making it difficult to extract information from datasets. The airborne particles analyzed in this paper are Sulphur dioxide (SO_2), Nitrogen Oxide (NO), Nitrogen Dioxide (NO_2), Carbon Monoxide and particulate matter (PM).

2.2. London's Sites

London is the largest urban area and capital city of the United Kingdom. Greater London covered an area of 1,579 square kilometers. A larger area, referred to as the London Metropolitan Region covered an area of 8,382. (Sumbler et al., 1996)

There are a number of monitoring sites that are available in London, England. For this work, only four sites were chosen due to the availability of the data for the five particles used in this research. These sites are: London Bexley, Bloomsbury, London Marylebone Road and London North

Kensington.

London Bexley's site is located about 13 meters above the ground in a suburban area around 200 meters from A206 Northend Rd. and 300mts from Thames Rd. London Bloomsbury site is located within a self-contained unit at the north-east corner of a central London gardens. All four sides of the gardens are surrounded by a busy 2 lane one-way road system, which is subject to frequent congestion. The nearest road lies at a distance of approximately 25 meters from the station. The manifold inlet is approximately 3 meters high. (Defra, 2009)

Furthermore, London Marylebone Road site is located in a self contained cabin on Marylebone Road opposite Madame Tussauds. The manifold inlet is located at a height of 3 meters from the ground. The nearest road, the A50 is approximately 1 meter from the station. Traffic flows of over 80,000 vehicles per day pass the site on six lanes. The road is frequently congested. Lastly, the site at London North Kensington is located within a self contained cabin in the grounds of Sion Manning School. The manifold inlet is approximately 3 meters from the ground. The nearest road is a quiet residential road approximately 10 meters from the station. The surrounding area is mainly residential (Defra, 2009).

2.3. Recurrence Plots

Recurrence Plot (RP) is a graphical tool introduced by Eckmann (1987) in order to extract qualitative characteristics of a time series. The recurrence of a state I at a different time j is pictured within a two-dimensional squared matrix with black and white dots, where the black dots represent a recurrence and both axes represent time (Zbilut et al., 1998; Aboofazeli, 2008).

Such RP can be mathematically expressed as:

$$R_{i,j}^{m,\varepsilon_i} = \Theta\left(\varepsilon_i - \left\|\vec{X_i} - \vec{X_j}\right\|\right), \ \vec{X_i} \ \mathbb{R}^m, \quad i,j = 1..N, \quad (1)$$

where, N is the number of considered states x_i; ε_i is a threshold distance, $\| \ . \ \|$ a norm and Θ (.) the Heaviside function (Furman, 2006).

Since $R_{i,i} = 1 \ (i = 1 \dots N)$ by definition, the RP has a black main diagonal line called line of identity (LOI). In this context, the Heaviside function Θ (.) is a recurrence of a state $\vec{X_i}$ that is sufficiently close to $\vec{X_i}$ (states $\vec{X_i}$ that fall into an m-dimensional neighborhood) (Bradley et al., 2002).

Using the time series of a single observable variable (particles, in this case), it is possible to reconstruct a phase space trajectory. Starting from the scalar time series $\{X_t\}_{t=1}^T$ a sequence of embedded vectors

$$y(i) \equiv \left(x_i, x_{i+\tau}, x_{i+2\tau}, \dots, x_{i+(m-1)\tau}\right)$$

is generated (Palmieri et al., 2009). The set of all embedded vectors $y_{(i)}, i = 1, \dots, T - (m-1)\tau$, constitutes a trajectory in $\mathfrak{R}^m$ where m is the embedding dimension and τ is the time delay. Each unknown point of the phase space at time I is reconstructed by the delayed vector $y(i)$ in an m-dimensional space called the reconstructed phase space.

According to several authors, determining the embedding parameters should be the first step for nonlinear analysis (Marwan, 2002; Palmieri et al., 2009; Gao et al., 2000; Aparicio, 2008). As recurrence plots are highly sensitive to several of the features mentioned previously; a small change in one of these parameters can change the appearance of recurrence plots significantly (Rohde et al., 2008). Therefore, a search for the best dimension and time delay must be made first. In this appraisal, the best dimension value is calculated using the algorithm of false nearest neighbors (FNN) as shown on (Zou, 2010; Palmieri et al., 2009).

Also, when calculating an RP a norm must be chosen [Karakasidis et al 2009]. The most widely used norms are the L1, L2 (Euclidean norm) and L∞ (Zbilut, 2002).

For this contribution, the Euclidean norm was used. Figure 1 shows the recurrence plots of a random signal, a sine wave and two RPs chosen randomly for airborne particle concentration.
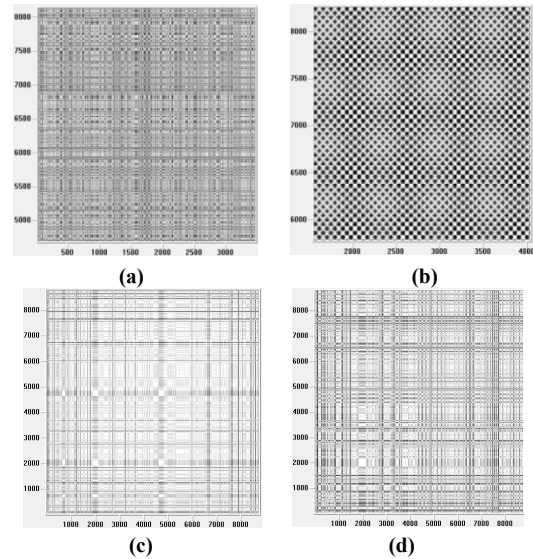
Figure 1. Recurrence Plots using (a) a random signal, (b) a sine wave, (c) particle concentration of carbon monoxide at London Bexley for 2010 and (d) particle concentration of sulphur dioxide over 2010 at Marylebone Road (daily mean) showing the Line of Identity (diagonal line).

Although it is possible to identify each plot from figure 1 (c and d), some experience is needed to interpret the RPs (Zbilut, 2007). For this reason, recurrence quantification analysis (RQA) offers a window to characterize RP structures.

The main idea of this project is to reconstruct the (unknown) system dynamics in the phase space by using time-delay embedding, and then computing the distances between all pairs of embedded vectors, generating a symmetric two-dimensional square matrix for each dataset as shown on figures 1c and 1d, applying RQA to each dataset.

2.4. Recurrence Quantification Analysis (RQA) for RPs

Zbilut (1998) and Webber (1994) have developed some of the methods used today for Quantitative Analysis of the recurrence plots. It has been shown that these measures are

able to capture dynamical transitions in complex systems (Zuo et al., 2010). They define measures of complexity using certain characteristics of the recurrence plots (March et al., 2005; Marwan, 2007).

In general, the characteristics measured in a RP are: recurrence rate, determinism, ratio, entropy and trend.

2.4.1. Recurrence Rate

The recurrence rate is a measure of recurrences, or density of recurrence points in the RP. This rate gives the mean probability of recurrences in the system (Marwan, 2007; Strozzi et al., 2007). The recurrence rate is given by:

$$RR(\varepsilon) = \frac{1}{N^2} \sum_{i,j=1}^{N} R_{i,j}(\varepsilon) \tag{2}$$

in the case of time series, and;

$$RR(\varepsilon) = \frac{1}{N^4} \sum_{i_1,i_2,j_1,j_2}^{N} R_{i_1,i_2,j_1,j_2}(\varepsilon) \tag{3}$$

in the case of spatial data [Mocenni et al, 2011].

The recurrence rate represents the fraction of recurrent points with respect to the total number of possible recurrences. It is a density measure of the RP.

2.4.2. Determinism

Determinism is a measure for predictability of the system (Aparicio, 2008). The determinism could also be explained as the percentage of recurrent points forming line segments which parallel the Line of Identity (LOI). The determinism characteristic is given by (Gao et al., 2000):

$$DET = \frac{\sum_{l=l_{min}}^{N} l P^{\varepsilon}(l)}{\sum_{i,j}^{N} R_{i,j}^{m,\varepsilon}} \tag{4}$$

Where P(l) denotes the probability of finding a diagonal line of length l in the RP. This measure quantifies the predictability of a system (Zou et al., 2010). The measure of determinism (DET) ranges from 0 to 1. Numbers near zero indicate randomness while those approaching one indicate the presence of a strong signal component (Furman et al., 2006; Ahlstrom, 2006).

The average diagonal line length L_{mean} is defined as:

$$L_{mean=} \frac{\sum_{l=l_{min}}^{N} l P(l)}{\sum_{l=l_{min}}^{N} P(l)} \tag{5}$$

This characterizes the average time that two segments of a trajectory stay in the vicinity of each other, and is related to the mean predictability time (Zou et al., 2010).

The choice of l_{min} can also be used in order to exclude short temporal scales that are not important. (Karakasidis, 2009)

2.4.3. Ratio

The Ratio variable is defined as the quotient of determin-

ism (DET) divided by the recurrence (REC). It is useful to detect transitions between states: this ratio increases during transitions but settles down when a new quasi-steady state is achieved (Palmieri et al., 2009).

2.4.4. Entropy

The measure characteristic entropy refers to the Shannon entropy of the frequency distribution of the diagonal line lengths (Yulmetyev et al., 1999). According to several authors, the basic idea is that information (Shannon) entropy of the random processes is abundantly supplied with the qualitative and quantitative data on the object under research (Marwan, 2002; Yulmetyev et al., 1999; Strozzi et al., 2007; Karakasidis et al., 2009). The entropy of a system is given by:

$$ENT = -\sum_{l=l_{min}}^{N} p(l) \log p(l) \; with \, p(l) = \frac{p^{\varepsilon}(l)}{\sum_{l=l_{min}}^{N} P^{\varepsilon}(l)} \tag{6}$$

2.4.5. Trend

The trend is a linear regression coefficient over the recurrence point density of the diagonals parallel to the LOI. The trend measurement is given by:

$$TREND = \frac{\sum_{i=1}^{N} \left(i - \frac{N}{2}\right)(RR_i - RR_i)}{\sum_{i-1}^{N} (i - \frac{N}{2})^2} \tag{7}$$

3. Experimental Results

Recurrence Quantification Analysis have been carried out for years 2005-2010 for all four sites mentioned in section 2.2 using the raw data (hourly) obtained from DEFRA (Defra, 2009) for each particle. The recurrence rate (REC), determinism (DET), Ratio, Entropy (ENT) and Trend have been modeled using Matlab® software. The results were analyzed separately and then put them together to present results altogether in form of boxplots. This analysis is complex due to the large quantity of the datasets.

There is much useful information that can be extracted from the recurrence plots using RQA. Figure 2 shows the recurrence rate for all particles.
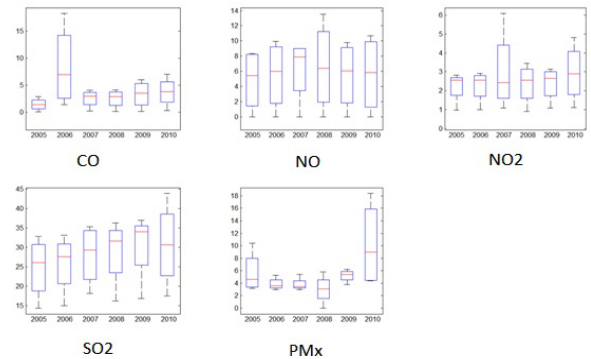
Figure 2.   Recurrence Rate for Particle Concentration, years 2005-2010.

In figure 2 is shown the recurrence Rate for all five particles (CO, NO, NO_2, SO_2 and PMx. In this figure, it is worth notice that the median recurrence rate for CO, NO NO_2 and PMx lies from 3 to 6, with the lowest recurrence rate being for nitrogen dioxide. However, Sulphur dioxide shows a much higher recurrence rate with an average of 29 increasing in some regions of London Bloomsbury to 44. This higher recurrence rate may be due to the low variances in values of the datasets for all years, making it easier for RQA to determine recurrence.
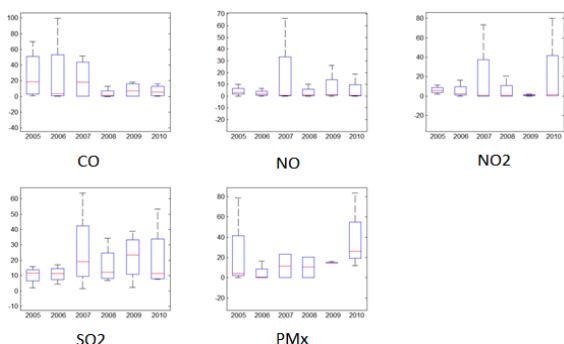

Figure 3.   Determinism for Particle Concentration, years 2005-2010.

Furthermore, the determinism for SO_2 is also higher than for other particles, having a median of 18 as shown on figure 3. Although it seems lower due to the scaling of the boxplots, the median shows otherwise, the spread in the 25th to 75th percentiles and the length of the whiskers may be due to exceptionally high determinism for that site in particular or an outlier and not necessarily represent a higher determinism altogether.
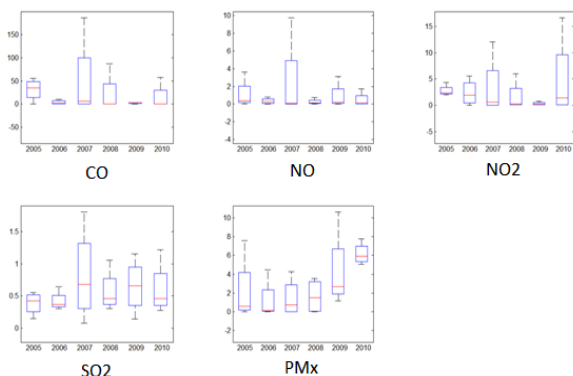

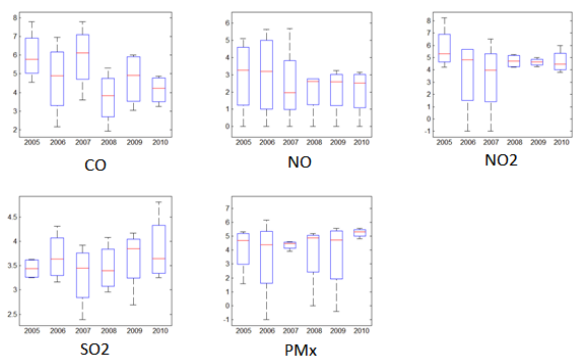**Figure 4.**   Ratio for Particle Concentration, years 2005-2010.

Figure 5.   Entropy for Particle Concentration, years 2005-2010.

Furthermore, it is worth notice for entropy that the frequency distribution of the data is slightly higher for particle concentration CO than for the other particles. The other particles seem to have steady entropy whose median oscillates between 3 to 5 with a few exceptions (i.e. PMx2006, 2009). This is shown on figure 5.

The last measure was the trend. Since the trend represents the measure of the positioning of recurrent points away from the central diagonal, that is the paling of the RP towards its edges (Palmieri et al., 2009). A ''flat'' diagram indicates stationarity, whereas drift in the signal will result in the overall increase or reduction of distances as the signal is moved away from the main diagonal. In this respect, it could be noticed that most of the particles have a median between -0.5 to 0.5. There are a few exceptions such as SO_2 for 2010 and PMx for 2010. The reason could not be ascertain for sure, hence further investigation is recommended. This is shown on figure 6.
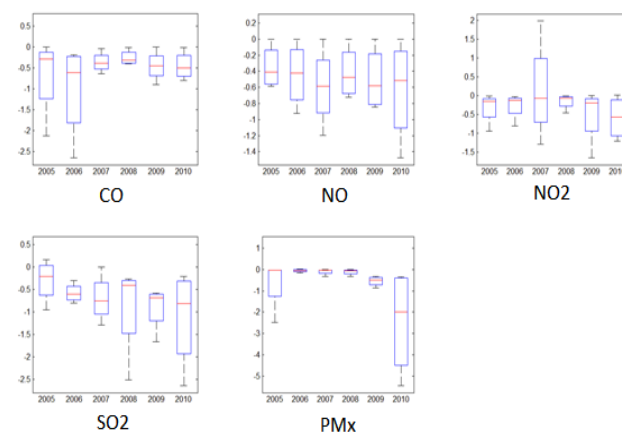

Figure 6.   Trend for Particle Concentration, years 2005-2010.

4. Conclusions and Future Work

Numerous experiments have been carried out with different particles and through different years. Using Recurrence Quantification Analysis it could be shown that information could be extracted from large datasets of dissimilar airborne particles during a considering time lap (six years, in this case). Trends could be identified using these tools and preliminary conclusions suggest that important information such as density distribution, drifts, among others could be drawn.

For future work, it could be useful to use a combination o RQA with prediction algorithms such as Support Vector Machines to carry out prognosis of the airborne particle data. Another useful approach that could be carried out is the use of cross recurrence plot (CRP), making a comparison between two recurrence plots to determine trends.

ACKNOWLEDGEMENTS

REFERENCES

[1] Aboofazeli M., Moussavi Z.K., Comparison of recurrence plot features of swallowing and breath sounds. Chaos, Solitons and Fractals. 37: 454–464, 2008

[2] Ahlstrom Christer, Höglund Katja, Hult Peter, Häggström Jens, Kvart Clarence, and Ask Per, Distinguishing Innocent Murmurs from Murmurs caused by Aortic Stenosis by Recurrence Quantification Analysis, World Academy of Science, Engineering and Technology. 18:40-45, 2006

[3] Aparicio T., Pozo E.F., Saura D., Detecting determinism using recurrence quantification analysis: Three test procedures. Journal of Economic Behavior & Organization. 65: 768–787, 2008

[4] Arbex M.A., Nascimento Saldiva P.H., Amador Pereira L.A., Ferreira Braga A.L., Impact of outdoor biomass air pollution on hypertension hospital admissions. J Epidemiol Community Health. 64: 573-579, 2010

[5] Bradley Elizabeth, Mantilla Ricardo, Recurrence plots and unstable periodic orbits. Chaos. 12- 3:596-600, 2002

[6] Defra (Department of Environment, Food and Rural Affairs), Air Pollution in the UK report: Latest Annual Report – 2009 Edition B, 2009

[7] Eckmann, J. P., Kamphorst, S. O. & Ruelle, D., Recurrence plots of dynamical systems. Europhys. Lett. 4:324–327, 1987.

[8] Furman Michael D., Simonotto Jennifer D., Beaver Thomas M., Using Recurrence Quantification Analysis Determinism for Noise Removal in Cardiac Optical Mapping. IEEE Transactions on biomedical Engineering. 53(4), 2006

[9] Gao Jianbo, Cai Huaqing., on the structures and quantification of recurrence plots. Physics Letters A. 270:75–87, 2000

[10] Guo Yuming, Barnett Adrian G, Zhang Yanshen, Tong Shilu, Yu Weiwei, Pan Xiaochuan, The short-term effect of air pollution on cardiovascular mortality in Tianjin, China: Comparison of time series and case–crossover analyses. Science of the Total Environment. 409:300–306, 2010

[11] Karakasidis T. E., Liakopoulos A., Fragkou A., Papanicolaou P., Recurrence Quantification Analysis of Temperature Fluctuations in a Horizontal Round Heated Turbulent Jet. International Journal of Bifurcation and Chaos. 19-8: 2487–2498, 2009

[12] Kilabuko J.H., Matsuki H., Nakai S., Air Quality and Acute Respiratory Illness in Biomass Fuel using homes in Bagamoyo, Tanzania. Int. J. Environ. Res. Public Health. 4(1): pp. 39-44, 2007

[13] Liu Yan-Ju, Harrison Roy M., Properties of coarse particles in the atmosphere of the United Kingdom, Atmospheric Environment. 45:3267-3276, 2011

[14] March T.K., Chapman S.C., Dendy R.O., Recurrence plot statistics and the effect of embedding. Physica D. 200: 171–184, 2005

[15] Marwan N., Thiel M., Nowaczyk N. R., Cross recurrence plot based synchronization of time series. Nonlinear Processes in Geophysics. 9:325–331, 2002

[16] Marwan Norbert, Romano M. Carmen, Thiel Marco, Kurths Jürgen, Recurrence plots for the analysis of complex systems. Physics Reports. 438:237 – 329

[17] Mirasgedis, S., Hontou, V., Georgopoulou, E., Sarafidis, Y., Gakis, N., Lalas, D. P., Environmental damage costs from airborne pollution of industrial activities in the greater Athens, Greece area and the resulting benefits from the introduction of BAT. Environmental Impact Assess Review. 28(1):39–56, 2008

[18] Mocenni Chiara, Facchini Angelo, Vicino Antonio, Comparison of recurrence quantification methods for the analysis of temporal and spatial chaos, Mathematical and Computer Modelling 53:1535–1545, 2011

[19] Palmieri Francesco, Fiore Ugo. A nonlinear, recurrence-based approach to traffic classification. Computer Networks. 53:761–773, 2009

[20] Rohde Gustavo K., Nichols Jonathan M., Dissinger Bryan M., Bucholtz Frank, Stochastic analysis of recurrence plots with applications to the detection of deterministic signals. Physica D. 237:619–629, 2008

[21] Strozzi Fernanda, Zaldívar Jose Manuel, Zbilut Joseph P., Recurrence quantification analysis and state space divergence reconstruction for financial time series analysis. Physica A. 376:487–499, 2007

[22] Sumbler M.G., *London and the Thames Valley*,British Regional Geology series, British Geological Survey, 1996

[23] Tanio Masaaki, Hirata Yoshito, Suzuki Hideyuki, Recurrence plot statistics and the effect of embedding, Physics Letters A 373:2031–2040, 2009

[24] Weinmayr G., Romeo E., De Sario M., Weiland S.K., Forastiere F., Short-Term Effects of PM10 and NO2 on Respiratory Health among Children with Asthma or Asthma-like Symptoms: A Systematic Review and Meta-Analysis. Circulation. 121:2331-2378, 2010

[25] Yang Hui,, Bukkapatnam Satish T.S., Barajas Leandro G., Local recurrence based performance prediction and prognostics in the nonlinear and nonstationary systems. Pattern Recognition. 44: 1834–1840, 2011

[26] Yin Jianxina, Harrison Roy M., Chen Qiang, Rutter Andrew, Schauer James J., Source apportionment of fine particles at urban background and rural sites in the UK atmosphere. Atmospheric Environment 44:841-851, 2010

[27] Yulmetyev R.M., Gafarov F.M., Dynamics of the information entropy in random processes. Physica A. 273:416-438, 1999

[28] Zbilut P., José-Manuel Zaldivar-Comenges, Fernanda Strozzi, Recurrence quantification based Liapunov exponents for monitoring divergence in experimental data. Physics Letters A. 297:173–181, 2007

[29] Zbilut J.P., Webber Jr. C.L., Recurrence Quantification Analysis: Introduction and Historical Context. International Journal of Bifurcation and Chaos. 17(10):3477–3481, 2007

[30] Zou Yong, Donner Reik V., Donges Jonathan F., Marwan Norbert, Kurths Jürgen, Identifying complex periodic windows in continuous-time dynamical systems using recurrence-based methods, Chaos. 20. 043130, 2010