From Nonlinearity to Causality: Statistical testing and inference of physical mechanisms underlying complex dynamics

Milan Paluš
Institute of Computer Science,
Academy of Sciences of the Czech Republic,
Pod vodárenskou věží 2,
182 07 Prague 8, Czech Republic
mp@cs.cas.cz; http://www.cs.cas.cz/mp
(Dated: January 12, 2008)

Principles and applications of statistical testing as a tool for inference of underlying mechanisms from experimental time series are discussed. The computational realizations of the test null hypothesis known as the surrogate data are introduced within the context of discerning nonlinear dynamics from noise, and discussed in examples of testing for nonlinearity in atmospheric dynamics, solar cycle and brain signals. The concept is further generalized for detection of directional interactions, or causality in bivariate time series.

Index

This is a preprint of an article whose final and definitive form has been published in Contemporary Physics 48(6) (2007) 307 – 348. **available online at** `http://www.informaworld.com/smpp/content~content=a792784686`

1. INTRODUCTION

One of the great challenges of contemporary physics and contemporary science in general is understanding of emergent phenomena in complex systems. Physicists try to apply their ideas and tools in studies of complex dynamics of various origin, thus it is not surprising that review papers in this journal discuss complex processes not only in traditional physical areas such as, for instance, the physics of atmosphere and climate [1, 2], but also touch complexity in biological systems in general [3], in proteins [4], DNA [5, 6], or in the dynamics of the human cardiovascular system [7]. Processing and evaluation of

complex biomedical signals and images is also a problem attracting the attention of physicists and the Contemporary Physics [8, 9]. Systems of many interacting components can be found not only in the traditional areas of statistical physics, but also at various levels of organization of living tissues and organisms, from molecular structures to neuronal networks, from proteins, cells, to the human brain, and further to interactions of human individuals or groups organizing themselves in social networks, working and doing business in the globalized economy. From some contemporary physicists we can hear the opinion that economics might be 'the next physical science' [10]. Others assert that the contemplation and resolution of questions at the interfaces of biology, mathematics, and physics promise to lead to a greater understanding of the natural world and to open new avenues for physics [11].

In the scientific areas close to biology or to social sciences, the connection between experimental data and a theory is usually less straightforward than in traditional physical sciences. The decision whether data support a proposed theory or a hypothesis usually cannot be made just using simple methods of data evaluation and presentation. Sophisticated statistical approaches should be used in order to distinguish repetitive patterns from random effects and then to infer possible physical mechanisms underlying an observed complex phenomenon. The traditional field of mathematical statistics provides both a language and a toolbox for dealing with the questions of inference that emerge in the search for 'order in chaos', or, more specifically, in attempts to discern noise from complex dynamics generated by possibly deterministic and probably nonlinear physical (chemical, biological, social) processes. A formal framework that can help in asking and answering questions about possible mechanisms underlying experimental data has been developed in the field of statistical testing. The statistical testing is widely used in many scientific fields including some areas of physics (see, e.g., [12–14, 16–18]), however, in many physical areas it is underestimated or ignored. The underestimation occurs in two ways: A part of physicists use it, however, without a deep understanding of the principles and without serious mastering of the available tools. This underestimation of the necessary expertise level leads to many incorrect results and/or wrong interpretations of the statistical tests, giving thus arguments to the proponents of the other kind of underestimation – the belief that the statistical testing as a scientific tool is useless (see, e.g., [19] and references therein). Therefore it is desirable to discuss realistic possibilities of the statistical testing in a form accessible to a broad physical readership.

This article is an attempt to introduce the statistical testing to a general physical community. Instead of starting with formal definitions and/or rephrasing materials from some of the many textbook of mathematical or applied statistics, we will briefly review the scientific development in the area of nonlinear dynamics and the theory of deterministic chaos, oriented to processing of experimental data, in which the need of the statistical testing naturally emerged.

A description and explanation of the ways of physical thinking that have led to complicated theories and their elegant formulations in the language of mathematics, instead of lecturing the chain of axioms, definitions, theorems and corollaries, was the way how Professor Jozef Kvasnica (1930–1992) used to give his memorable lectures on quantum mechanics and quantum field theory at the Faculty of Mathematics and Physics of the Charles University in Prague. The author would like to dedicate this article to *memory of Professor Jozef Kvasnica*, head of department of mathematical physics (1976–1986), a theoretical physicist and a great educator [20] who was able to maintain a relatively liberal academic atmosphere at his department in 1980s Czechoslovakia.

In experimental studies of complex systems, usually it is not possible to characterize the state of such systems by a single measurement or a set of measurements. In many cases, however, it is possible to follow dynamics or evolution of a system by recording of some observable quantity (or a set quantities, or the same quantity in different spatial locations) by registering its values in successive instants of time $t_1, t_2, \ldots, t_N$. The collection of measurements $\{s(t)\}$ is called a *time series*. The time series are the kind of data we will consider in this article. Traditional statistical approaches to time series analysis are based on the linear theory [21], although some extensions counting for specific types of nonlinearity have been developed [22, 23]. An independent approach to analysis of nonlinear time series emerged in physics-related areas of nonlinear dynamics and the theory of deterministic chaos. We will introduce some ideas of this approach below. Inspirations from the chaos theory have been explored in the statistical context of time series analysis as well [23, 24].

Having recorded a time series $\{s(t)\}$ of some observable quantity reflecting complex behaviour of some system or process, one could ask questions about the nature of that system or process. Is it random, or is it deterministic and predictable? If it is not random, what are the underlying mechanisms? Can we infer them from the data and express them in a form of a mathematical model with a predictive power? These questions are quite general questions of science. We will describe how researchers in the field of nonlinear dynamics and chaos tried to cope with such questions and how the statistical testing has been introduced into the 'search for order in chaos'. After an introduction of necessary concepts of statistical testing, namely the computational Monte Carlo approaches based on the surrogate data techniques, we will discuss the problem of detection of nonlinearity in time series. Starting with some theoretical considerations, we will continue our review with particular exam-

ples of testing for nonlinearity in atmospheric dynamics, solar cycle and in brain signals. From statistical inferences about dynamics of a single process we will move to a study of interactions between systems, in particular, we will demonstrate a statistical approach for detection of directional interactions, or causality from bivariate time series.

2. SEARCHING ORDER IN CHAOS

In 1991 Shu-yu Zhang published the Bibliography on Chaos [25] comprising 7167 titles from which more than 2700 papers contained the words 'chaos', 'chaotic', or 'strange attractor' in their title.

The editor of the New York Times Book Review chose as one of the best books of the year 1987 James Gleick's Chaos: Making a New Science [26]. Gleick [27] characterized the 'new science of chaos' as the scientific revolution shifting the paradigm of understanding natural and social phenomena. Indeed, the 'science of chaos', or, to be more precise, the theory of deterministic chaos [28] has found its way not only to physics and other natural sciences, but also to philosophy [29, 30] and to the art [31]. The chaos theory has been proposed as a new principle for a successful management of firms and corporations [32]. What has made the chaos theory so attractive for scientists, journalists, philosophers and artists?

The first important feature of chaotic dynamics is that it can give a simple explanation to complex behaviour. For instance, the following two equations, known as the Hénon system [33]

$$\begin{aligned} x_{n+1} &= 1.4 - x_n^2 + b\, y_n \\ y_{n+1} &= x_n \end{aligned} \qquad (1)$$

represent a deterministic description for a movement in a plane. Each point has two coordinates, x and y, which are evolving in discrete time instants $n = 1, 2, \ldots$. The transition from the point (x_n, y_n) to the following point (x_{n+1}, y_{n+1}) is exactly given by the prescription according to the equations (1). For $b = -0.01$, after some transient time, the Hénon system remains forever jumping between two points. The temporal evolution of the coordinate x (illustrated as the plot of x_n vs. n) shows a simple periodic pattern (the top panel in Fig. 1). For $b = -0.001$, the after-transient state consists of eight different points. The temporal evolution of coordinate x is slightly more complicated, but still shows repeating patterns (the middle panel in Fig. 1). For $b = 0.3$, even the after-transient state consists of an infinite number of different points, although confined to a bounded subset of the plane. In the temporal evolution of coordinate x some patterns may resemble the previous ones, however, no pattern exactly repeats (the bottom panel in Fig. 1.) This is an example of the deterministic non-periodic behaviour, called the chaotic dynamics. Does the term 'deterministic' mean that such dynamics is predictable? Let

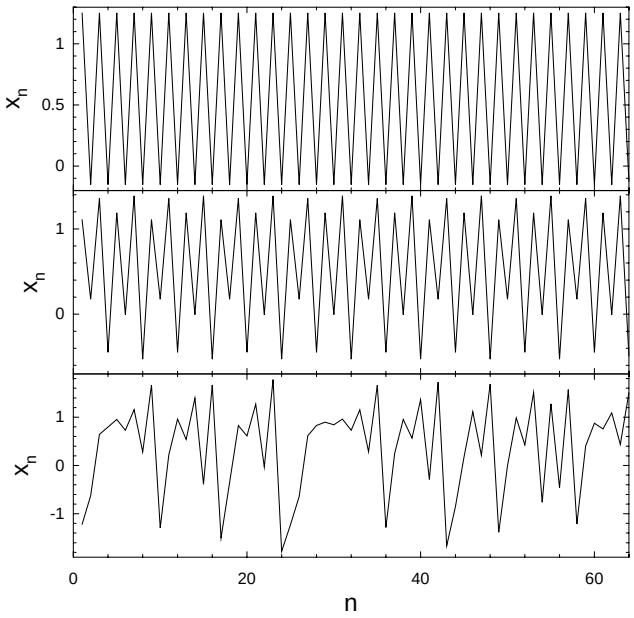

FIG. 1: Temporal evolution of coordinate x of the Hénon system for $b = -0.001$ (top panel), $b = -0.01$ (middle panel) and for the chaotic state $b = 0.3$ (bottom panel).

us consider two temporal evolutions starting at nearby points (x_0, y_0) and $(x_0 + \Delta x, y_0)$. Further, let us consider $\Delta x = 0.2$ and the periodic state of the Hénon system with $b = -0.001$. The two trajectories starting in the two points (x_0, y_0) and $(x_0 + \Delta x, y_0)$ evolve close to each other (the top panel in Fig. 2). If we take $\Delta x = 0.02$, the two trajectories are even undistinguishable in the precision of our graphics (the middle panel in Fig. 2). With the small $\Delta x = 0.02$, but in the case of the chaotic Hénon system with $b = 0.3$, the two trajectories are close to each other only during a few iterations and then they evolve as two different trajectories (the bottom panel in Fig. 2). This behaviour is called the sensitive dependence on initial conditions, or the exponential divergence of trajectories and has fatal consequences for the predictability of chaotic systems. Even if we know the exact equations for a particular chaotic system, for predicting the future we need to measure the present state. Each measurement, however, has some error. In the above example we can consider Δx as the measurement error and the initial points (x_0, y_0) and $(x_0 + \Delta x, y_0)$ as the actual state and the state given by the measurement with the finite precision. Then the two trajectories can be considered as the actual one (the solid line) and the predicted one (the dashed line in Fig. 2). We can see that the measurement error propagates into the prediction error of a comparable magnitude in the case of the regular (periodic) state of the system. In the chaotic state, however, the prediction error grows in an exponential rate. The two trajectories in the bottom panel in Fig. 2 do not exponentially move away each other simply because the dynamics of the system is confined within a bounded set in a plane. The

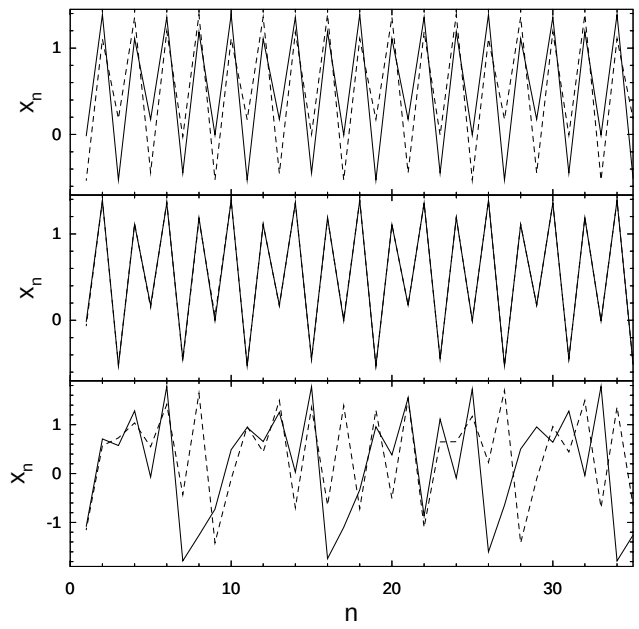

FIG. 2: Temporal evolution of coordinate x of two trajectories (solid and dashed lines) starting at two nearby points with distance $\Delta x = 0.2$ in the case of $b = -0.01$ (top panel); $\Delta x = 0.02$ and $b = -0.01$ (middle panel); and $\Delta x = 0.02$ in the chaotic state with $b = 0.3$ (bottom panel).

prediction error, however, very soon reaches the magnitude comparable with the diameter of this set, so that any prediction is impossible.

Now we know that we can generate complex dynamical behaviour with limited or practically none predictability by simple model equations. Can such models, however, help to understand natural complex phenomena? A usual way, applicable to regular (non-chaotic) phenomena would be fitting a chosen mathematical model using the experimental data and then testing the validity of the model by comparing the predictions, given by the model, with the measured data. Considering the limited predictability of the chaotic dynamics, however, this approach is hardly applicable.

A breakthrough in the chaos theory and its connection to the real world was the seminal paper 'Geometry from a time series' [34]. In this paper, Packard et al. [34] demonstrated how to reconstruct a subset of abstract mathematical space in which variables of a chaotic system evolve (such as the above variables x and y in the equations (1)) using just one measurable quantity reflecting information about the system state. Recording of an observable quantity of a system under study in the form of a time series can thus be a key to uncover the hidden order in the observed chaos. Having reconstructed the 'geometry of behaviour', it is then possible to estimate a dimension of such a state space and thus to know how many equations are needed to explain the observed complex dynamics. This way of characterization of low-dimensional chaotic systems (i.e. systems described by n ordinary differential equations, when n is sufficiently low) based on the processing of experimental time series has also been understood as a way to detect chaotic dynamics underlying complex experimental data. The arguments were simple: Recorded complex experimental time series reflects dynamics which cannot be explained by any regular deterministic model. In the dimensional analysis we obtained a low dimension n, so we observe a process generated by an n-dimensional chaotic deterministic system, since noise cannot be explained by a finite-dimensional process. These considerations will be more precisely formulated in the following Sections 2 2.1 and 2 2.2.

Early enthusiasm of many researchers has led to straightforward applications of this approach and to claiming experimental evidence for the existence of (low-dimensional) deterministic chaos in many types of systems and processes, ranging from atmospheric dynamics [35–37], solar cycle [38–40], through complex physiological phenomena such as the dynamics of the heart [44, 45] and brain activity [46–48], to complex social phenomena such as the financial markets [41–43], just to mention a few examples.

2.1. Low-dimensional chaotic dynamics: detection and characterization

Let us consider that possible states of a system can be represented by points in a finite dimensional state space, say $\mathcal{R}^d$ (also the term phase space is used). The transition from the system's state $\mathbf{x}(t_1)$ at time t_1 to its state at time t_2 is then given by a deterministic rule: $\mathbf{x}(t_2) = T_{t_2-t_1}(\mathbf{x}(t_1))$. This transformation can be expressed either in continuous time by a set of ordinary differential equations:

$$\dot{\mathbf{x}}(t) = \mathbf{F}(\mathbf{x}(t))\,, \tag{2}$$

or in discrete time $t = n\Delta t$ by a map of $\mathcal{R}^d$ onto itself:

$$\mathbf{x}_{n+1} = \mathbf{f}(\mathbf{x}_n)\,. \tag{3}$$

The transition rules T_t in the forms (2) or (3), are called a *dynamical system*. Various forms of $\mathbf{F}$ (resp. $\mathbf{f}$) lead to different types of dynamics, ranging from simple dynamics of fixed points and limit cycles to complex, irregular behaviour. If the dynamics is dissipative (contracting its 'vital space', i.e. the subset of the state space visited by the system's trajectory), the area visited by the system after some transient time will be concentrated on a subset of the phase space. This set is called an *attractor*. Since not all points on an attractor are visited with the same frequency, it is possible to define a measure $\mu(\mathbf{x})d\mathbf{x}$, as the average fraction of time which a typical trajectory spends in the phase space element $d\mathbf{x}$. In an *ergodic* system, the measure $\mu(\mathbf{x})$ is the same for almost all initial conditions. Averages over phase space taken with respect to $\mu(\mathbf{x})d\mathbf{x}$ are then equal to time averages taken over a typical trajectory.

Let us consider now that a registered time series $\{s(t)\}$ is a scalar projection of a trajectory $\{\mathbf{x}(t)\}$ of a dynamical system in $\mathcal{R}^d$, i.e. $\{s(t) = s(\mathbf{x}(t))\}$. In other words, there is a d-dimensional dynamical system underlying the studied process, however, the full d-dimensional trajectory $\{\mathbf{x}(t)\}$ cannot be observed. An observer has an access just to a scalar observable $s(t)$, registered as the time series $\{s(t)\}$. Following the original idea of Packard et al. [34], a reconstruction of the original trajectory can be obtained using so-called time-delay embedding

$$\mathbf{s}(t) = (s(t-(m-1)\tau), s(t-(m-2)\tau), \ldots, s(t))\,, \quad (4)$$

where τ is the *delay time* and m is the *embedding dimension*. Provided that the measurement function $s : \mathcal{R}^d \to \mathcal{R}$ is sufficiently smooth (at least C^2) and that either the dynamics or the measurement function is generic in the sense that it couples all degrees of freedom, then for $m \geq 2d+1$, Takens [49] proved that the reconstructed trajectory is topologically equivalent to the original one. This problem is discussed and the 'embedding theorem' further generalized in [50].

The topological equivalence of the reconstructed and the original trajectories means that characteristics of the 'geometry of dynamics', i.e. of the geometry of an attractor of the dynamical system computed from the trajectory reconstructed by the time-delay embedding, are the same as if they were estimated from the original trajectory, or, ideally, from the invariant measure $\mu(\mathbf{x})d\mathbf{x}$ generated by the original trajectory. This way of a possible connection between complex, noisy-like experimental data and a mathematical chaotic model is the second feature which made the chaos theory so attractive for researchers. This new way of processing experimental data has sometimes been considered as an alternative to standard statistical methods of time series analysis.

The dynamics of chaotic systems is characterized by the above-mentioned exponential divergence of near-by trajectories. On the other hand, when the trajectory after the transient state evolves on an attractor, it is confined to a limited subset of the phase space. This combination of stretching, folding, and volume contraction lead to irregular geometry of the system's attractor which can have a statistically self-similar structure. This fractal geometry of the system's attractor can be characterized by fractal dimensions. Several definitions of non-integer dimensions have been proposed in the literature [28], e.g. the well-known Hausdorff dimension, or the computationally more accessible box counting dimension. One can weight points in a set by the average frequency with which they are visited by the trajectory. Then a definition of the dimension in terms of the natural measure $\mu(\mathbf{x})d\mathbf{x}$ on the attractor can be obtained. Attempts to take weighted averages of the number of points contained in elements of a partition of the phase space and study their dependence on the refinement of the partition usually did not bring satisfactory results due to limited amounts of available data. An alternative way to define the dimension of a measure $\mu(\mathbf{x})d\mathbf{x}$ was proposed by Grassberger and Procaccia [51]. They estimate the locally averaged density ρ_ϵ as the convolution of μ with a kernel function $K_\epsilon(r) = K(r/\epsilon)$ of bandwidth ϵ falling off quickly enough so that the convolution exists:

$$\rho_\epsilon(\mathbf{x}) = \int_{\mathbf{y}} d\mathbf{y}\mu(\mathbf{y})K_\epsilon(\|\mathbf{x}-\mathbf{y}\|)\,. \quad (5)$$

Usually, the kernel is chosen to be $K_\epsilon(r) = \Theta(1-r/\epsilon)$ where $\Theta(\cdot)$ is the Heaviside step function, $\Theta(x) = 0$ if $x \leq 0$ and $\Theta(x) = 1$ for $x > 0$. Then the correlation integral of order q is given as:

$$C_q(\epsilon) = \int_{\mathbf{x}} d\mathbf{x}\,\mu(\mathbf{x})\,[\rho_\epsilon(x)]^{q-1}\,. \quad (6)$$

For a self-similar measure, for $\epsilon \to 0$

$$C_q(\epsilon) \propto \epsilon^{(q-1)D_q}\,. \quad (7)$$

D_q is called the dimension of order q. This definition includes the dimension D_0 which is equivalent to the Hausdorff dimension in many cases. Grassberger and Procaccia [51] proposed the *correlation dimension* D_2 as a means of quantifying the 'strangeness' of an attractor. The term *strange attractor* was coined in the chaos literature for attractors of dynamical systems which were chaotic, i.e. their dynamics was characterized by the exponential divergence of trajectories and the geometry of the underlying attractor was fractal. The fractality is captured by any non-integer fractal dimension, such as the correlation dimension D_2. The chaoticity, i.e., the exponential divergence of trajectories is characterized by the positive Lyapunov exponents. Let us briefly review basic ideas of so-called fixed evolution time algorithm for estimating the largest Lyapunov exponent according to Wolf et al. [52].

Given a scalar time series $x(t)$, an m-dimensional trajectory is reconstructed using the time-delay method [49] according to Eq. (4), i.e., $\mathbf{x}(t) = \{x(t), x(t+\tau), \ldots, x(t+[m-1]\tau)\}$, where τ is the delay time and m is the embedding dimension. A neighbour point $\mathbf{x}(t')$ is located so that the initial distance δ_I, $\delta_I = ||\mathbf{x}(t) - \mathbf{x}(t')||$, is $s_{\min} \leq \delta_I \leq s_{\max}$, where $||.||$ means the Euclidean distance. The minimum and maximum scales $s_{\min}$ and $s_{\max}$, respectively, are chosen so that the points $\mathbf{x}(t)$ and $\mathbf{x}(t')$ are considered to be in a common 'infinitesimal' neighborhood. After an evolution time $T \in \{1, 2, 3, \ldots\}$, the resulting final distance δ_F is calculated: $\delta_F = ||\mathbf{x}(t+T) - \mathbf{x}(t'+T)||$. Then the local exponential growth rate per a time unit is:

$$\lambda_1^{\text{local}} \;=\; \frac{1}{T}\log(\delta_F/\delta_I). \quad (8)$$

To estimate the overall growth rate, in the case of deterministic dynamical systems the largest Lyapunov exponent (LLE) λ_1, the local growth rates are averaged along the trajectory:

$$\lambda_1 \;=\; \langle\lambda_1^{\text{local}}\rangle \;=\; \frac{1}{T}[\langle\log(\delta_F)\rangle - \langle\log(\delta_I)\rangle], \quad (9)$$

where $\langle . \rangle$ denotes averaging over all initial point pairs fulfilling the condition $s_{\min} \leq \delta_I \leq s_{\max}$.

The algorithm of Wolf et al. [52] quickly became one of the most frequently used methods for detecting chaos in experimental time series (after the Grassberger-Procaccia correlation dimension algorithm [51]), although later several methods for estimating the full Lyapunov spectrum, i.e., n Lyapunov exponents of an n-dimensional dynamical systems, characterizing the exponential divergence of trajectories in each of the n directions, have been developed [53–55].

2.2. Ubiquity of chaos challenged: surrogate data testing

Various types of dimensions were used to characterize the 'geometry of behaviour', i.e., the dimensionality and fractality of attractors of underlying dynamical systems. The 'chaoticity' of dynamics, i.e., the sensitive dependence on initial conditions can be characterized by Lyapunov exponents and dynamical entropies such as the Kolmogorov-Sinai entropy [28, 56, 57]. Methods of estimating dimensions and entropies of chaotic systems have been discussed in several review papers, see e.g. [58–61], as well as in books [62, 63]. Also a number of conference proceedings are devoted to this topic, for instance Refs. [56, 64–66]. A critical review of nonlinear dynamics and chaos related methods for time series analysis has been written by Schreiber [67]. The book of Kantz and Schreiber [63] was followed by a useful program package described also in Ref. [68].

Among the 'measures of chaos', i.e., dimensions, entropies and Lyapunov exponents, the correlation dimension D_2 [51] became the most frequently used quantity estimated from experimental time series in order to detect and characterize a strange attractor underlying analysed data. Having a measured time series $\{s(t)\}$, in the typical approach the time-delay embedding (4) was constructed for an increasing sequence $m_1, m_2, m_3, \ldots$ of embedding dimensions m. For each m, the correlation dimension $D_2(m)$ was estimated. Usually, the $D_2(m)$ estimates were increasing with increasing m, but from some m_s the estimates of $D_2(m)$ 'saturated', i.e. a sequence

$$D_2(m_1) < D_2(m_2) < \cdots < D_2(m_s - 1) < D_2(m_s) \approx D_2(m_s + 1) \approx D_2(m_s + 2) \approx \ldots$$

was obtained. Then the estimate $D_2(m_s)$ was considered as the fractal dimension of the underlying attractor and m_s as the sufficient embedding dimension and as the number of equations necessary to explain the observed dynamics, reflected in the recorded time series $\{s(t)\}$. Such an 'evidence for low-dimensional chaos' has been observed in time series obtained from very different sources already mentioned above. Here we only remind the data and processes which we will discuss later: The sunspot numbers representing the solar cycle [38–40], time series of meteorological variables reflecting the atmospheric dynamics [35–37], and human EEG – electroencephalogram, recordings of electrical potentials of the brain [46–48]. The broad usage of the correlation dimension for detection of chaos evoked also critical studies which pointed out some limitations of the method [69–71]. It was shown that with a limited amount of data the estimator of the correlation dimension saturated on a finite (and low) value even when so-called white noise (realizations of independent, identically distributed random variables) was used as the input data. Krakovská [72] derived an analytical formula quantifying underestimation of the correlation dimension given the number of data samples and the embedding dimension. Statistical estimators for fractal dimensions and their theoretical properties are also studied in Refs. [73–77].

Understanding the effect of the time series length, i.e. of the number of available data samples on the estimation of the correlation dimension (as well as on other dimensions) has led to the first 'test of reality' of a low-dimensional chaos. Such a test was realized as a comparison of the estimation of the correlation dimension from studied data with results of the same estimator applied to the same number of samples of white noise. White noise, however, similarly as a purely deterministic system, is a mathematical construct rarely occurring in the nature. If a time series $\{s(t)\}$ is a realization of a white-noise process, then any two samples, say $s(t)$ and $s(t + \tau)$ are independent. In natural processes subsequent values $s(t)$ and $s(t + \tau)$ can be dependent. A class of linear stochastic processes characterized by power-law spectra exhibit long-range linear dependence. Realizations of such processes, also known as coloured (pink, or red) noise, were analysed by Osborne and Provenzale [78]. They demonstrated that applications of the standard Grassberger-Procaccia algorithm yielded a finite correlation dimension for time series generated by stochastic systems with power-law spectra [78]. This problem is also discussed by Theiler [79].

A serial dependence between $s(t)$ and $s(t + \tau)$ in the form of the linear autocorrelation is a property consistent with an explanation by a linear stochastic process, i.e., a sequence of random variables, or, in other words, noise that is not white. The linear autocorrelation be-

tween $s(t)$ and $s(t+\tau)$ can be the source of a spurious finite dimension estimation. Various corrections have been proposed [67, 73] for dealing with this problem, as well as with other sources of bias in dimension estimations [80, 81]. Some authors tried to remove the linear dependence present in the data by fitting a linear regression model and using the model residuals in further analysis. Such an approach is established in statistics when searching for dependence structures beyond the linear correlations. Theiler & Eubank [82], however, demonstrated that this procedure was not suitable for the detection of chaotic dynamics, since the removal of the linear dependence destroyed structures of an underlying attractor. In time series from nonlinear dynamical systems, dependence structures cannot usually be linearly decomposed into a sum of linear and nonlinear parts and analysed separately. As a way how to assess the validity of finite dimension estimations, Theiler et al. [83] proposed to construct a special kind of filtered noise, i.e. control data as realizations of a linear stochastic process which possesses the same linear properties as the data under study. As the linear properties, namely the autocorrelation function $c(\tau)$ is meant, giving the linear dependence between $s(t)$ and $s(t+\tau)$ for the time series $\{s(t)\}$. Such test data can be obtained when a linear stochastic process with the same spectrum as the analysed data is constructed. An application of the dimensional estimates to such a 'control data set', in order to ascertain that the finite dimension estimates were not a numerical artifact, has been proposed by several authors [38, 83–86], however, Theiler et al. [83] introduced a systematic approach using the statistical language of hypothesis testing. For the control data sets Theiler et al. [83] coined the term *surrogate data*.

The surrogate data with the same sample spectrum as the tested time series can be constructed using the fast Fourier transform (FFT). The FFT of the series is computed, the magnitudes of the (complex) Fourier coefficients are kept unchanged, but their phases are randomized. The surrogate series is then obtained by computing the inverse transform into the time domain. Different realizations of the process are obtained using different sets of the random Fourier phases.

Having constructed a set of independent surrogate data realizations, the correlation dimension (or other 'nonlinear' measure) is estimated from the surrogate time series in order to obtain a range of values typical for a linear stochastic process with the given spectrum and the autocorrelation function. Then the value of the correlation dimension (or other quantity) obtained from the analysed experimental data is compared with the surrogate range, asking the question 'is the tested data set (time series) significantly different from the related surrogate data?' If the answer is 'yes', then the possible explanation of the data by a linear stochastic process is rejected and the result is considered as an evidence for nonlinearity, the necessary condition for the deterministic chaos. If the answer is 'no', then, in the statistical language, the so-called null hypothesis of a linear stochastic process is accepted. This decision, however, does not mean that an evidence for an explanation of the data by a linear stochastic process was obtained. Such a result merely means that the used measure, e.g., the correlation dimension, is not able to distinguish the analysed data from the isospectral linear stochastic process and the hypothesis about an underlying low-dimensional chaotic system is unfounded, even if the absolute value of the estimated correlation dimension is low. Below we will discuss these issues in detail.

3. STATISTICAL TESTING

The mathematical science of statistics is focused to the collection, analysis, interpretation or explanation, and presentation of data. Methods of *descriptive statistics* are used to describe a collection of data. Structures or patterns in the data may be modelled, accounting for randomness and uncertainty in the observations. Using the statistical models, inferences are drawn about processes underlying the data, or about the population from which the data were sampled; this is called *inferential statistics*. Although statistical methods have been developing for centuries, according to experts in the field [87], the modern inferential statistic began with Karl Pearson's work [88] on the goodness-of-fit test, motivated by the question whether an assumed probability model adequately describes the analysed data. Pearson's paper [88] is regarded as one of the twenty most important scientific breakthroughs of the 20th century along with advances and discoveries like the theory of relativity, the IQ test, hybrid corn, antibiotics, television, the transistor and the computer [89]. During the 20th century, the modern inferential statistics developed from its infancy into a mature science and nowadays it is applied in almost all scientific disciplines, as well as in technology, business and decision-making.

Considering such a broad usage, it is not surprising that also criticism of the statistical testing appeared in various scientific fields [19, 90–92]. Detailed analyses show, however, that the critical voices are valid as far as they concern numerous incorrect applications and interpretations of the statistical tests. One of the problem leading to the major mistakes in interpreting the results of the statistical tests is the mixing of two different testing paradigms – the Fisher's significance testing and the Neyman-Pearson's hypothesis testing. Therefore we need a distinctive definitions of the two approaches. The third – Bayesian approach will not be discussed in this paper.

3.1. Significance testing and hypothesis testing

A statistical test is a framework for searching answers to questions about experimental data. As the data, let us consider a (set of) measurement(s) x of a random vari-

able (process) X. A property of the data under interest should be quantifiable by some real function $\theta(x)$, called a *discriminating statistic*, or simply a *statistic*. We set a *null hypothesis* H_0: $\theta = \theta_0$. For simplicity, let us consider $\theta_0 = 0$ and $\theta(x) \geq 0$. Due to the random nature of X, there is a nontrivial probability function for $\theta(x) \geq 0$ even if the null hypothesis $\theta = 0$ is valid. Consider that we know or can calculate this probability function for the specified null hypothesis.

In the framework of the *Fisher's significance testing* [93–95] we ask what is the probability that we can observe an outcome of magnitude $\theta(x)$ (or larger) providing that the null hypothesis $\theta = 0$ is valid. We compute *significance*, or p-value as $p = \text{Prob}(\theta(x^0) \geq \theta(x))$, where x^0 are realizations of the random variable (process) X under the condition of the validity of H_0. The p-value is a measure of inductive evidence against H_0, smaller the p-value, the stronger the evidence. Small p-value indicates an unlikely event and, hence, an unlikely hypothesis. Fisher saw statistics as playing a vital part in inductive inference, drawing conclusions from the particular to the general, from samples to a population, or from a measurement to an underlying process. Scientific knowledge is created via inductive inference, for Fisher the evidential p-value had an important role in this process. The evidential p-value is a data-dependent random variable.

The *Neyman-Pearson's hypothesis testing* [96] is not a theory of statistical inference, but a paradigm for using the statistical testing as a mechanism for making decisions and guiding behaviour. While Fisher specified only the null hypothesis H_0, Neyman & Pearson introduced two hypotheses, the null H_0 and the alternative H_1 and their approach requires a decision between two distinct courses of action, accepting H_0 or rejecting it in favour of H_1. When choosing between accepting H_0 and H_1 mistakes occur. The significance level in the Neyman-Pearson's approach, usually denoted as α and called the Type I error, is the probability of the false rejection of H_0 (a false positive result). The Type II error, β, is the probability of the false acceptance of H_0 (a false negative result). The Neyman-Pearson's approach is not concerned with gathering evidence, it is aimed at error minimization. The validity of errors is understood in a long run, considering the *frequentist principle:* In repeated practical use of a statistical procedure, the long-run average actual errors should not be greater than (and ideally should equal) the long-run average reported error [97]. The nominal significance level α is chosen before the test is done. From the equation $\alpha = \text{Prob}(\theta(x^0) \geq \gamma)$ the critical value γ is calculated and the null H_0 is always rejected if $\theta(x) > \gamma$ is obtained for the tested data x.

3.2. Null distribution of discriminating statistic

In the classical approach to significance and hypothesis testing, a discriminating statistic θ is carefully tailored to match the null hypothesis. Statistics are preferred which have standard or 'standardized' distributions for a given null hypothesis, i.e., the distribution of θ for the specified null hypothesis can be analytically derived and the critical value γ for a given significance level α can be easily calculated using a known formula or found in a table. Tables of critical values for the most frequently used discriminating statistics and null hypotheses can be found in many statistical handbooks, see e.g. Refs. [98–100]. The restriction of the statistical testing to the statistics with known distributions has to some extent been superseded by computer-intensive methods, in which the distribution of θ and its confidence range for a given α can be accurately estimated by so-called Monte-Carlo simulation. Such an approach has been proposed by Barnard [101] and fully developed by Hope [102] and others [103–114]. The basic idea of this approach is to compute values of θ for many different realizations of the null hypothesis and to empirically estimate the distribution of θ from these values. Considering the quickly increasing performance and availability of powerful computers, Efron in his 1979 manifesto [108] argued for replacing (not always possible) analytical derivations based on (not always realistic) narrow assumptions by computational estimation of empirical distributions for statistics under interest. In computational statistics, the term *bootstrap* [112] is coined for Monte-Carlo methods and random resampling of experimental data, frequently with the aim to estimate confidence intervals ('error bars') for evaluated statistical quantities [108–112]. We should be careful to distinguish the quite different problems of estimating confidence intervals and testing null hypotheses, although the resampling, bootstrap or surrogate data methods can be technically similar. In the problem of the confidence interval estimation, a statistic of some intrinsic interest is computed from the data, and the goal is to find the error bars using realizations of resampled data obtained from a (probability) model aimed at the true underlying distribution. In the case of hypothesis testing, there is a specific, carefully stated null hypothesis, and the goal is to test whether the data are consistent with that hypothesis. The latter is the case of the surrogate data technique [83] proposed for the inductive inference of nonlinearity in the studied data by constructing the surrogate data as realizations of a linear stochastic process replicating the 'linear properties' of the original data. We propose to extend the term surrogate data for general significance testing using statistics and null hypotheses for which an analytical treatment is not known or not possible. As a working definition we can say that the surrogate data are numerically generated data which preserve all important statistical properties of the original data but the property which is tested for. More realistic definition for practical purposes requires that the surrogate data replicate those data properties which influence the discriminating statistic and are the primary source of its bias and variance irrespectively of the presence of the tested property. Of course, the surrogate data should not posses the tested

property. We will discuss these issues using concrete examples in the following Sections 5 and 6.

3.3. The story of sunspots and senators as an illustrative example

The energy output of the Sun as the main basis of life on the Earth is nearly constant. However, the Sun is far from being uniform. The best observed solar inhomogeneities are spots on the solar surface in which the luminosity is diminished but magnetic fields appear which are stronger than usual magnetic fields on the rest of the solar surface.

In the middle of the 19th century it was discovered by a druggist H. Schwabe and described in the scientific literature by Wolf [115] that the number of spots on the Sun varies in a cyclic manner with a characteristic time of about 11 years. Initially, this variation was considered periodic, but nowadays its aperiodicity and complex behaviour is stressed and studied in detail. We will return to the problem of the intrinsic dynamics of the solar cycle in Sec. 6.

The historical data of the sunspot numbers have been attracting researchers since the 1852 Wolf's paper [115]. Recently, in the very intensive debate about the global warming and its possible causes, the sunspot numbers, as the long-term record of variability of the solar activity, are studied and compared with various climate-related data [116–118]. On the other hand, some climatologists oppose that no clear evidence exists for the solar variability causing the climate change, and the most famous declared correlation between the parameters of the solar cycle and the global atmospheric temperature [119] was obtained by incorrect processing and evaluating of the data [120, 121]. Although the cause of the climate change is an important problem which is very interesting also from the viewpoint of statistical testing, it is beyond the scope of this article. In the following we will only borrow an example used by some climatologists [122] as well as statisticians [123] in order to demonstrate pitfalls in testing and interpretation of correlations.

The top panel of Figure 3 shows the number of the Republicans in the U.S. Senate in the years 1960 – 2006 [124], while the middle panel shows the sunspot numbers in the same period [125]. Let us consider N observations x_i, y_i of two variables x and y. For assessing the linear dependence of the variables x and y by the means of (cross)correlation c(x,y), it is necessary to 'normalize' the variables by subtracting their means

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

and dividing by their standard deviations σ

$$\sigma^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2$$

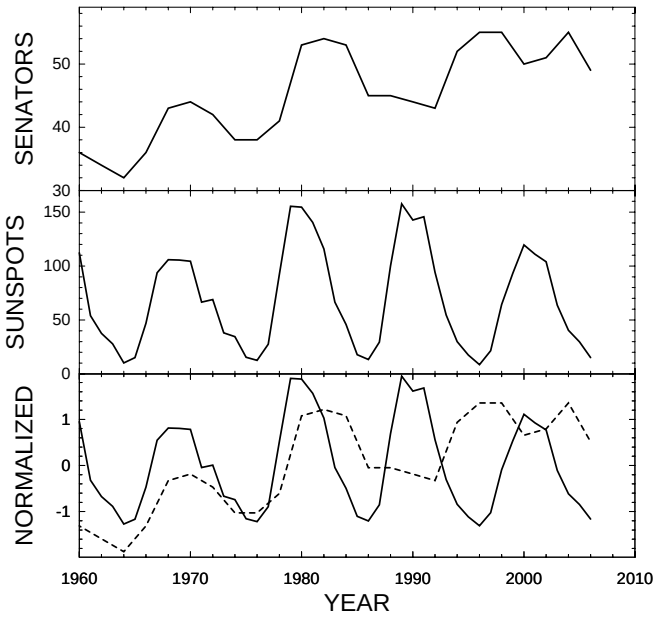

FIG. 3: The number of Republicans in the U.S. Senate in the years 1960 – 2006 (top panel); the sunspot number for the period 1960 – 2006 (middle panel); the number of the Republican senators (dashed line) and the sunspots (solid line) normalized to zero mean and unit variance (bottom panel).

in order to obtain normalized

$$\widetilde{x_i} = \frac{x_i - \bar{x}}{\sigma}\,.$$

The same procedure is done with the measurements y_i in order to obtain normalized $\widetilde{y_i}$. Then, the correlation between x and y is

$$c(x,y) = \frac{1}{N}\sum_{i=1}^{N}\widetilde{x_i}\widetilde{y_i}\,.$$

The normalized numbers of the sunspots and the Republican senators are plotted in a common scale in the bottom panel of Fig. 3. We can see that in the first part of the record, from 1960 to 1986, the agreement between the numbers of the sunspots and the senators is remarkably good. The correlation coefficient c obtained from the related 14 samples (a part of the Senate is elected each second year, so the number of the senators is sampled biannually) is $c = 0.52$. In the classical statistical testing the correlation (coefficient) is a very common statistic and its distribution for the null hypothesis of linear independence (H_0: $c = 0$) is well known for data sets sampled from independent, identically distributed, normally distributed populations (IID Gaussian distributions). For such a null hypothesis the critical values γ for given significance levels α can be easily computed or find in tables in statistical handbooks. The correlation coefficient attains values from -1 to 1, so it can differ from the zero value either in the negative or in the positive direction.

Thus in general one should test the digression from the zero value in both the directions and the 'two-tailed' or 'two-sided' test should be applied. If we are interested only in the positive correlation (such as the one observed between the numbers of sunspots and senators in the period 1960–1986), we apply the one-sided (one-tailed test). Choosing the significance level $\alpha = 0.05$, the related critical value $\gamma = 0.458$ says that 95% of the distribution of the correlation coefficient under the condition of validity of the above specified null hypothesis is bounded by $c \leq 0.458$, considering the one-sided test and $N = 14$ samples. In the statistical tables the term 12 degrees of freedom (df) is used, considering df $= N - 2$. For $N = 24$ samples (here the whole record 1960–2006) the critical value is $\gamma = 0.344$. For the period 1960–1986 the correlation $c = 0.52 > \gamma = 0.458$, i.e. the null hypothesis should be rejected. Supposing that only the 1960–1986 data are available, a possible temptation to interpret this result as an evidence for a causal relation between the solar activity and the election success of the Republicans is used by some authors as a demonstration of a very limited value of statistical testing for inferring dependence when a physical mechanism of interactions in not known. They argue further that this apparent nonsense can be statistically disproved only when more data is available – using the whole record 1960–2006 the null hypothesis of independence cannot be rejected ($c = 0.16 < \gamma = 0.344$).

In fact, this argument against the value of the statistical testing for inference of new knowledge in climatology, and in physical sciences in general, is based on an incorrect interpretation of an incorrectly applied statistical test. Not only the Neyman-Pearson's approach is not suitable in this case, more importantly, the chosen null hypothesis is inadequate. In the following we will demonstrate that with an appropriately constructed surrogate data reflecting an adequate null hypothesis, the observed correlation $c = 0.52$ is not statistically significant, since it could probably occur by chance in the case of independent processes of adequate statistical properties.

For a better understanding of the approach let us start with the Monte-Carlo computational realization of the above inadequate null hypothesis that both the numbers of senators and sunspots were drawn from IID Gaussian distributions. For each analysed segment 1960–1986 (14 samples) and 1960–2006 (24 samples) we randomly draw independent pairs of 14- and 24-sample sets and compute their correlations $c(x, y)$. We draw 100,000 independent realizations of the null hypothesis and then estimate distributions of $c(x, y)$ as 100-bin histograms from the obtained 100,000 values. These histograms are presented by solid lines in the upper panels of Fig. 4 and Fig. 5 for the 24- and the 14-sample sets, respectively. If, instead of the two IID Gaussian draws, we correlate the numbers of the senators on the one side (variable x) with the random Gaussian draws as the variable y on the other side, the resulted histograms are practically the same (dashed lines almost coinciding with the solid lines in upper panels of Figs. 4, 5). Practically the same histograms are

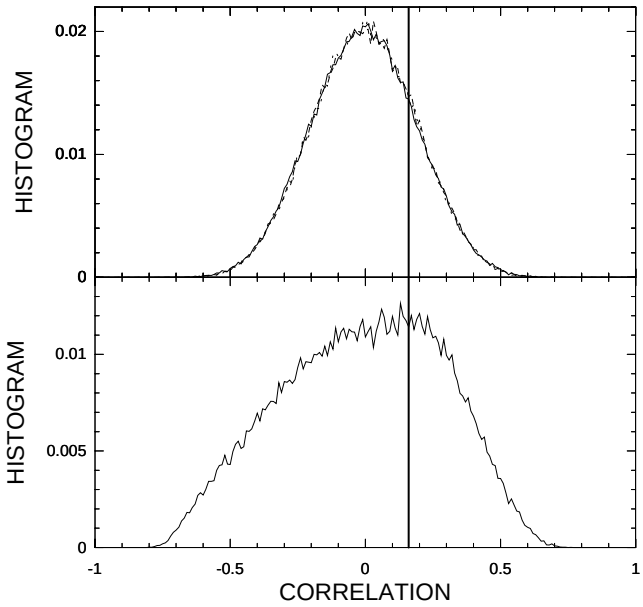

FIG. 4: Top panel: Distribution (histogram) of correlations between two 24-sample sets randomly drawn from a Gaussian distribution (solid lines); the histogram of correlations between the number of the Republican senators in the period 1960–2006 (24 samples) with 24-sample sets randomly drawn from the Gaussian distribution (dashed line); and the histogram of correlations between the number of the Republican senators in the period 1960–2006 (24 samples) with the related 24-sample segment of the sunspot numbers randomly permutated in the temporal order (IID surrogate, dash-and-dotted line). Bottom panel: Histogram of correlations between the number of the Republican senators in the period 1960–2006 (24 samples) with 24-sample realizations of the Barnes model. The vertical solid line shows the correlation between the number of the Republican senators and the sunspot numbers for the period 1960–2006.

again obtained when we correlate the numbers of the senators with the related numbers of the sunspots which are, however, permutated in the temporal order (dash-and-dotted lines in upper panels of Figs. 4, 5, mostly coinciding with the solid lines). The latter Monte-Carlo realizations of the null hypothesis (the sunspot numbers randomly permutated in temporal order; different realizations of the null hypothesis are obtained as different independent permutations), i.e. computationally generated data obtained by numerical manipulation of the original data, are usually called the surrogate data. The term surrogate data is also used for data generated by a model fitted on the original data. The terminology, however, is not unified. Some authors would also use the term 'surrogate data' for the above draws from the Gaussian distributions, while others would identify the surrogate data generation process as a bootstrap. The IID or 'scrambled' surrogate data are realizations of white noise (IID process) with exactly the same sample histogram as the original data. Of course, such data are by construction also independent from other data or processes, in

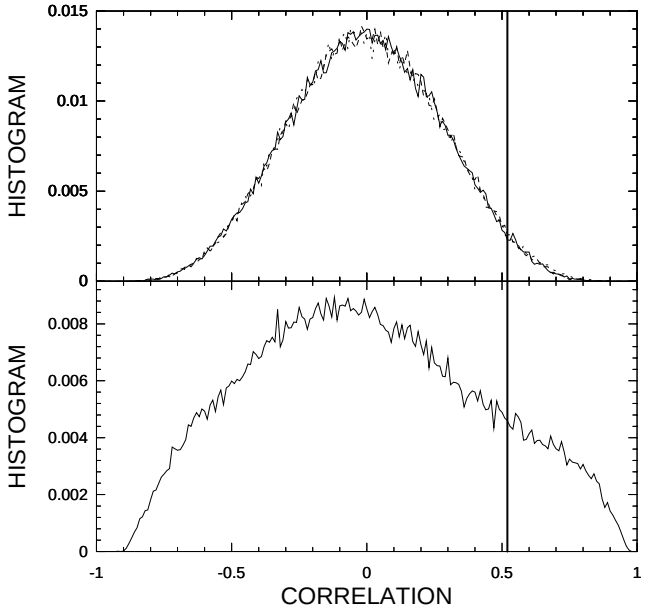

FIG. 5: Top panel: Distribution (histogram) of correlations between two 14-sample sets randomly drawn from a Gaussian distribution (solid lines); the histogram of correlations between the number of the Republican senators in the period 1960–1986 (14 samples) with 14-sample sets randomly drawn from the Gaussian distribution (dashed line); and the histogram of correlations between the number of the Republican senators in the period 1960–1986 (14 samples) with the related 14-sample segment of the sunspot numbers randomly permutated in the temporal order (IID surrogate, dash-and-dotted line). Bottom panel: Histogram of correlations between the number of the Republican senators in the period 1960–1986 (14 samples) with 14-sample realizations of the Barnes model. The vertical solid line shows the correlation between the number of the Republican senators and the sunspot numbers for the period 1960–1986.

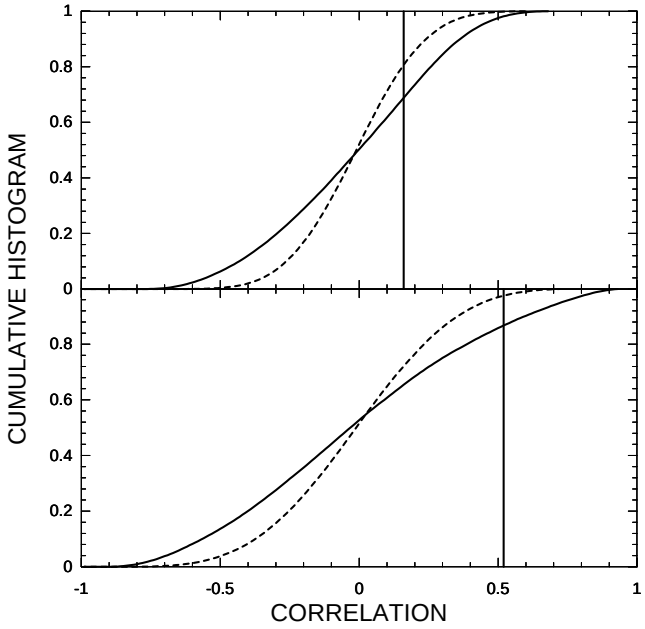

FIG. 6: Top panel: Cumulative histogram for the distribution of the correlations of the 24-sample sets drawn from the Gaussian distribution (dashed line); cumulative histogram for the distribution of the correlations of the number of the Republican senators in the period 1960–2006 (24 samples) with 24-sample realizations of the Barnes model (solid line). The vertical solid line shows the correlation between the number of the Republican senators and the sunspot numbers for the period 1960–2006. Bottom panel: The same as in the top panel, but for the 14 samples in the period 1960–1986.

this case from the number of the Republican senators.

In order to evaluate the statistical significance in the one-sided test, the histograms are summed-up from the most left bin to the right, in order to obtain the cumulative histograms (Fig. 6) showing the cumulative probability of correlation values from the minimum (-1) to a given value. It means that the significance p of a given value, i.e. the probability that a correlation equal to or larger than a given value could occur by chance, if the null hypothesis is valid, is the complement to one from the actual cumulative probability. For the 24-sample segment 1960–2006 we have $c = 0.16$ and we obtain $p < 0.22$ (cf. the vertical solid line with the dashed curve in the upper panel of Fig. 6). From the same cumulative histogram (the dashed curve in the upper panel of Fig. 6) we can read the critical value for the significance level $\alpha = 0.05$ which is, in the 24-sample case, $\gamma = 0.34$. The Monte-Carlo approach is consistent with the standard 'significance table' approach, advising not to reject the null hypothesis in the case of the record 1960–2006. For the sub-segment 1960–1986 again the Monte-Carlo agrees with the related tables: the significance for $c = 0.52$ is $p < 0.027$ (cf. the vertical solid line with the dashed curve in the lower panel of Fig. 6), the result in favour for rejecting the null hypothesis. Should we consider this result as an indication for a possible relation between the solar activity and the election success of the Republican Party? No. So, what is wrong? In fact, nothing. Let us remind the null hypothesis, stated above: independent ($c = 0$) realizations of IID Gaussian processes. As we demonstrated by using the IID surrogates, even the strict Gaussianity is not the decisive property, but the explanation of the failure of the current null hypothesis is in the IID property. The independent, identically distributed samples mean that in the data there is no relation between any x_i and x_{i+j}. Apparently, this is not true. The subsequent values x_i and x_{i+j} (or y_i and y_{i+j}) in both the data sets are not independent. Especially the sunspot numbers with their cyclic behaviour are characterized by non-zero autocorrelations $c(x_i,\ x_{i+j})$. This phenomenon, also called serial correlation, is typical especially in the context of time series and explains why the IID null hypothesis is inadequate in cases when $c(x_i,\ x_{i+j})$ and $c(y_i,\ y_{i+j})$ are nonzero for $j \neq 0$. The serial correlation is a data-specific property and therefore no universal table of critical values cannot be derived for testing the independence of serially correlated data sets,

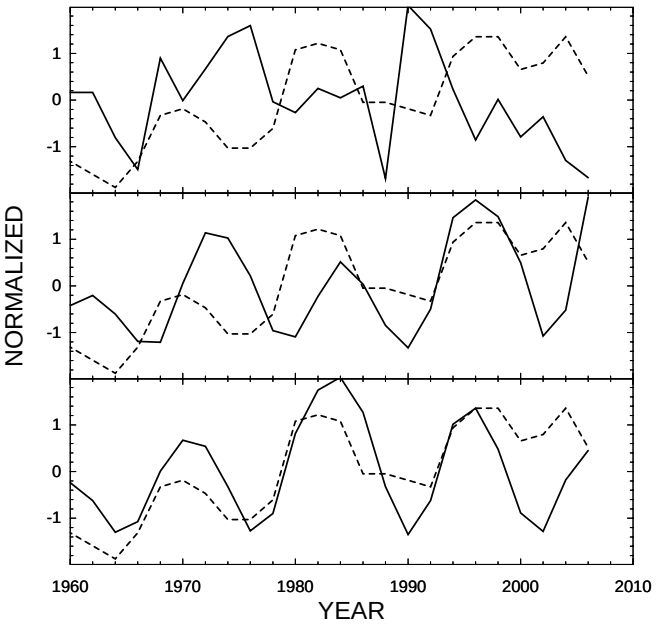

FIG. 7: Normalized numbers of the Republican senators (dashed line in all three panels) compared with random surrogate data (solid lines) obtained as a realization of Gaussian white noise (top panel) and two different realizations of the Barnes model for the sunspot cycle (middle and bottom panels).

or time series. Computing correlations and comparing them with tabelated critical values implicitly means that the IID null hypothesis is used.

The possible solution for testing the null hypothesis of independence for serially correlated data is the surrogate data approach. Here we will keep the senators data unchanged, but we construct surrogate data having the same serial correlations, or more precisely, the same autocorrelation function as the original sunspot data, however, by the construction the surrogate data are independent from the senators data, as well as from the original sunspot data. Such surrogate data can be either obtained by numerical manipulation of the original sunspot data, or generated by a mathematical model mimicking statistical properties of the sunspot numbers. We will use realizations of the Barnes model [126] which will be described in detail in Sec. 6, where we also will see that the linear dependence given by the autocorrelation function of the realizations of the Barnes model is consistent with the linear dependence in the sunspot data. Here we only demonstrate two different realizations of the Barnes model in Fig. 7. There is no structure in the realizations of the IID Gaussian process (the top panel of Fig. 7, solid line), while the Barnes model (the middle and bottom panels of Fig. 7) produces data with the same periodicity as that observed in the sunspot cycle. The artificial sunspot cycle, simulated by the Barnes model, is, as a process, independent from the senators data by the way of its constructions. A majority of realizations of the Barnes model have no apparent correlations with the senators data in the period 1960–1986, such as the example in Fig. 7, the middle panel. By chance, however, strongly correlated realizations can occur (Fig. 7, the bottom panel). Of course, the probability of such a chance requires a statistical evaluation.

Again, using 100,000 realizations of the Barnes model we construct the histograms of the correlations $c(x, y)$ with the senators data for the 24-sample and the 14-sample sets (the lower panels in Figs. 4 and 5, respectively). We can see that these histograms are wider than the related histograms for the IID null hypothesis, indicating that larger correlation values can occur by chance than in the case of the IID data. The cumulative histograms for the Barnes model surrogate data are depicted in Fig. 6 by the solid lines. The significance for the correlation value $c = 0.52$ for the period 1960–1986 is now $p < 0.14$ which is not considered small enough for rejecting the null hypothesis of independence. Of course, the correlation $c = 0.16$ for the period 1960–2006 is not significant either ($p < 0.32$). If we use the cumulative histograms obtained using the Barnes model surrogate data (Fig. 6, the solid lines) for evaluation of the test critical values, for the significance level $\alpha = 0.05$ for the one-sided test we have $\gamma = 0.45$ for 24 samples, and $\gamma = 0.73$ for the 14-sample series. Thus in the latter case, for the correlation to be significant on $\alpha = 0.05$, the correlation should exceed the value 0.73. The actual value $c = 0.52$ is not significant. The statistical test with the appropriate null hypothesis indicates that this correlation probably occurred by chance and campaign managers of the Republican Party need not to be afraid of a low solar activity.

Using a properly chosen surrogate data for numerical realizations of the null hypothesis in the Monte Carlo approach, the statistical testing can provide a powerful tool for preventing false positive detections of various phenomena and relationships. However, the choice of a correct testing methodology is not a trivial problem and the surrogate data should be carefully tailored for a particular case. In the following Sections we will discuss some problems encountered in surrogate data applications which recently appeared in a part of the physical literature.

Looking back at the $c(x, y)$ histograms obtained using the Barnes model surrogate data (the lower panels in Figs. 4 and 5) we can already at this point give a notice to a common error in the surrogate data applications. Many authors evaluate surrogate tests by computing a difference between the discriminating statistic obtained from the tested data and the mean value of the discriminating statistic obtained from the surrogate ensemble, divided by the standard deviation (SD thereafter) of the discriminating statistic obtained from the surrogate ensemble. Then, if such a difference is greater than 1.699, the difference is considered significant and the null hypothesis is rejected. Again, we do not go to subtle disputes between the Fisher's and Neyman-Pearson's

approaches, we rather remind that this critical value is derived from a Gaussian distribution of the discriminating statistic, where the digression 1.699SD is related to the significance level $\alpha = 0.05$ for rejecting the null hypothesis of a zero value in the two-sided test. We can see that the distributions of the correlations for the Barnes model surrogates are quite different from Gaussian ones. It is typical for various discriminating statistics, used in nonlinear dynamics in combinations with some specific surrogate data that the distributions of the statistics are wider than the Gaussian distribution and/or are fat-tailed (decreasing in a power-law way instead of the exponential decay as in the case of the Gaussian distribution). Therefore, it is recommended to evaluate the surrogate data test by estimating the exact significance p from the empirical distribution, rather than using the critical values derived from the Gaussian distribution, although the former requires much more realizations of the surrogate data. However, not all tests evaluated by using the 'number of SD' approach are incorrect. If the null hypothesis was not rejected for the difference smaller that 1.699SD, or, on the other hand, the null was rejected by a huge difference in higher multiples or tens of SD, then the correct histogram approach probably would not change the results. The test results in the range 1.7 – 3SD might be problematic and require a reassessment.

4. TESTING FOR NONLINEARITY

The motivation for and introduction of the surrogate data tests for nonlinearity into nonlinear dynamics and the theory of deterministic chaos has been described in Sec. 2 2.2. We can see this problem from a broader perspective of analysis of time series recorded in complex systems. Many authors have the temptation to apply modern, just developed, or even still being developed nonlinear methods in an expectation to obtain more information than by using the classical time series analysis methods [22], based on the linear theory. Various systems include nonlinear components so that a linear description seems unsatisfactory. The question is, however, whether a specific signal recorded from such a system is indeed affected by nonlinearity so that its linear analysis can fail. A nonlinearity test should be performed before using nonlinear methods which are usually more computer-intensive and theoretically/methodologically (mathematically/statistically) less elaborated than the linear methods. As we point out in a more general discussion on the statistical testing, failure to reject the null hypothesis of a linear process is not an evidence that the underlying mechanism is indeed linear. When the analysed data, however, cannot be, in a properly done test, distinguished from realizations of a linear process, than we can hardly expect that useful information can be extracted by a nonlinear descriptor. Moreover, there is a danger of unfounded interpretations of the results in usually so attractive and philosophically tractable terms of dimensions or Lyapunov exponents.

4.1. Formulation of the problem

Let $\{y(t)\}$ be a time series, i.e., a series of measurements done on a system in consecutive instants of time $t = 1, 2, \ldots$. The time series $\{y(t)\}$ can be considered as a realization of a stationary linear stochastic process $\{Y(t)\}$. Without loss of generality we can set its mean to zero. Then the linear stochastic process $Y(t)$ can be written as:

$$Y(t) = Y(0) + \sum_{i=1}^{\infty} a(i)Y(t-i) + \sum_{i=0}^{\infty} b(i)N(t-i)\,, \quad (10)$$

where $b(0) = 1$, $\sum_{i=1}^{\infty} |a(i)| < \infty$, $\sum_{i=0}^{\infty} |b(i)| < \infty$, $\{N(t)\}$ is an independent, identically distributed (iid), normally distributed process with zero mean and finite (constant) variance. (For more details see [21].)

Alternatively, the time series $\{y(t)\}$ can be considered as a projected trajectory of a dynamical system, evolving in some measurable d-dimensional state space. To be more specific, let X_t denote a state vector in R^d. Then the measurements $y(t)$ are obtained as $y(t) = g(X_t)$, where $g(.)$ is a projection (measurement function), and temporal evolution of X_t may be described by a discrete-time dynamical system (a difference equation):

$$X_t = F(X_{t-1})\,, \quad (11)$$

with $X_0 \in R^d$ and for $t \geq 1$.

Due to ubiquity of noise, it is more realistic to replace the above states by random variables and the dynamics by a Markovian model such as

$$X_t = F(X_{t-1}, e_t)\,, \quad (12)$$

where $t \in Z_+$, $F : R^{2d} \to R^d$, $\{e_t\}$ is a sequence of independent and identically distributed d-dimensional random vectors and e_t is independent of X_s, $0 \leq s < t$. We call $\{e_t\}$ the *dynamic noise.* Following Tong [24], we refer to equation (11) as the *skeleton* of model (12), considering $F(X) = F(X, 0)$. For convenience, it is frequently assumed that the dynamic noise is additive so that equation (12) reduces to the *model with additive noise*

$$X_t = F(X_{t-1}) + e_t\,. \quad (13)$$

4.2. The null hypothesis and the surrogate data

The approach for identifying nonlinearity in time series, introduced by Theiler et al. [83] as the method of surrogate data, seeks an evidence for nonlinearity by rejecting the null hypothesis of a linear stochastic process such as the above definition (10). It is important that the

Monte Carlo realizations of the null hypothesis – the surrogate data preserve the 'linear properties' of the tested data. It is meant, namely, the serial linear correlations, for time series $\{y(t), t = 1 \dots N\}$ expressed as the autocorrelation function

$$c(\tau) \equiv c(y(t), y(t-\tau)) = \frac{1}{N-\tau} \sum_{t=\tau+1}^{N} y(t)y(t-\tau)\,. \quad (14)$$

Without a loss of generality we consider that the time series $\{y(t)\}$ has a zero mean and a unit variance. According to the Wiener-Khinchin theorem the autocorrelation function is related to the power spectrum which can be estimated as the *periodogram* $P(k)$, using the Fourier transform of $\{y(t)\}$,

$$S(k) = \sum_{t=0}^{N-1} y(t)e^{i2\pi kt/N} \quad (15)$$

as $P(k) = |S(k)^2|$.

For a Gaussian linear process, the linear properties are specified by the squared amplitudes of the Fourier transform (i.e., the periodogram estimator of the power spectrum). The required surrogate time series $\{y'(t)\}$ can be generated by multiplying the Fourier transform of the data by random phases and then transforming back into the time domain:

$$y'(t) = \sum_{k=0}^{N-1} e^{i\alpha_k} \sqrt{P(k)} e^{-i2\pi kt/N}\,, \quad (16)$$

where $0 \leq \alpha_k < 2\pi$ are independent, uniformly distributed random numbers. Due to the way of generating the surrogate data using the Fourier transform (FT), we call this type of surrogate data the *FT surrogates*.

In order to capture an autocorrelation function of an underlying process, one can directly fit onto data a model of the same type as the definition of a linear process (10), just of a finite order m, n, i.e. the first sum in (10) has m summands and the second sum has n summands. Such a model is called an autoregressive moving average model, or an ARMA model, with specification of its order shortly written as ARMA(m,n). In principle, such a model can also be used for generating the surrogate data. It is necessary, however, to realize that we distinguish two types of null hypotheses - simple and composite. An example of a simple null hypothesis is a linear Gaussian process with a given mean and variance. In nonlinearity tests, however, we should consider a class of such processes without a specified mean and variance. Such a specification defines the composite null hypothesis. Parameters, specifying particular members of the considered class of processes are called the *nuisance parameters*. A discriminating statistic that does not depend on the nuisance parameters is called *pivotal*. Using a non-pivotal statistic to test a composite null hypothesis, when nuisance parameters are unknown, might be problematic.

Theiler & Prichard [127] explain that ARMA models produce *typical* realizations of a Gaussian process, while the FT surrogates are *constrained* realizations – constrained to a particular sample spectrum. Thus the FT surrogates provide a subset of all realizations that are typical for the considered process. Theiler & Prichard [127] demonstrate that for non-pivotal statistics and composite null hypotheses, the constrained surrogate data have better test performance than the typical realizations generated by the ARMA models. Although some nonlinear statistics are pivotal [128], the FT surrogates became most frequently used, probably also due to the elegant way of their generation. The performance of the FT surrogates in nonlinearity tests, however, can be jeopardized by pitfalls on different levels: problems of principal as well as of technical nature can be identified and not always satisfactorily solved. Mammen and Nandi [129] warn that the variance of critical values of a statistic obtained from the FT surrogates may be of the same order as the variance of the test statistic itself. Chan [130] shows that asymptotically the FT surrogates are exactly valid for stationary Gaussian circular processes. The FT surrogates tend to preserve the periodic autocorrelation function [67] rather then the standard autocorrelation function (14). This problem and possible solutions are discussed in Refs. [67, 131]. Paluš [132] analyses a related problem of strongly cyclic data, when a limited series length causes 'blurring' of a sharp spectral peak so that the FT surrogate data have weaker autocorrelation than the original data. This property can lead to false rejections of the null hypothesis. Therefore Paluš [132] proposes to accompany each nonlinearity test with a test using a linear statistic assessing the quality of the surrogate data.

Realizations of the FT surrogates tend to have a Gaussian distribution. A possibly different distribution of the original data can be another source of false positive results of nonlinearity test. Therefore Theiler et al. [83] propose a histogram transformation (rescaling of the data) in order to equal the histograms of the original data and the surrogates, then called the latter the amplitude adjusted FT (AAFT) surrogates. Using such a histogram transformation, in fact, one reformulates the basic null hypothesis of a linear stochastic process with Gaussian innovations (10) (and a Gaussian distribution) to a null hypothesis of a linear stochastic process (10), realization of which underwent a static nonlinear transformation, i.e., if $\{y(t)\}$ is a realization of (10), we measure $\{w(t)\}$ such that $w(t) = G[y(t)]$, where G is a nonlinear function. Such nonlinearity is called the static nonlinearity, since it is caused by a static transformation, e.g., it occurred during a measurement, but it is not inherent to the dynamics of the process.

It is shown in Ref. [133], however, that for short and strongly correlated sequences, the AAFT algorithm can yield an incorrect test since it introduces a bias towards a slightly flatter spectrum. In fact, the formal requirement for the surrogates to have the same sample periodogram

as the tested data should be extended by requiring also the same sample probability distribution (amplitude histogram) as the data. Schreiber and Schmitz [133] propose a method which iteratively corrects deviations in the spectrum and in the distribution. Alternatingly, the surrogate is filtered towards the correct Fourier amplitudes and rank-ordered to the correct distribution. The accuracy of such iterative amplitude-adjusted (IAFT) surrogates depends on the size and structure of the data. Considering these complication, a nonlinear statistic that is independent of the univariate data distribution can be very useful. Such a statistic used by Paluš [132] will be introduced in the following Sec. 4 4.3. A detailed discussion of the problems encountered in the generation of the surrogate data within the problem of the testing for nonlinearity was written by Schreiber and Schmitz [134].

4.3. Information-theoretic functionals as discriminating statistics

Information theory [135, 136] provides efficient mathematical tools for quantifying information carried by random variables and stochastic processes, information transmitted between and among variables and processes, as well as common information contained in two or more variables or processes. The quantification of common information led to measures of general dependence, i.e., the dependence beyond the linear correlation. Therefore information-theoretic measures have found their way to nonlinear dynamics and related nonlinear time series analysis methods. Here we will define some basic terms, necessary for further considerations.

Consider n discrete random variables $X_1, \ldots, X_n$ with sets of values $\Xi_1, \ldots, \Xi_n$, respectively. The probability distribution function (PDF) for an individual X_i is $p(x_i) = \Pr\{X_i = x_i\}$, $x_i \in \Xi_i$. We denote the PDF by $p(x_i)$, rather than $p_{X_i}(x_i)$, for convenience. Analogously, the joint PDF for the n variables $X_1, \ldots, X_n$ is $p(x_1, \ldots, x_n)$. The *redundancy* $R(X_1; \ldots; X_n)$, in the case of two variables also known as the *mutual information* $I(X_1; X_2)$, quantifies average amount of common information, contained in the n variables $X_1, \ldots, X_n$:

$$R(X_1; \ldots; X_n) \; = \;$$

$$\sum_{x_1 \in \Xi_1} \cdots \sum_{x_n \in \Xi_n} p(x_1, \ldots, x_n) \log \frac{p(x_1, \ldots, x_n)}{p(x_1) \ldots p(x_n)} \,. \quad (17)$$

When the discrete variables $X_1, \ldots, X_n$ are obtained from continuous variables on a continuous probability space, then the redundancies depend on a partition ξ chosen to discretize the space. Various strategies have been proposed to define an optimal partition for estimating redundancies of continuous variables (see [65, 132, 137] and references therein). Here we use a simple box-counting algorithm based on marginal equiquantization [57, 132, 138], i.e., a partition is generated adaptively in one dimension (for each variable) so that the marginal bins become equiprobable. It means that the marginal boxes are not defined equidistantly but so that there is approximately the same number of data points in each marginal bin. The only parameter of this method is the number Q of the marginal bins. Paluš [132] proposed that computing redundancy/mutual information of n variables, the number of marginal bins should not exceed the $n+1$-st root of the number of the data samples, i.e. $Q \leq \sqrt[n+1]{N}$.

The equiquantization method effectively transforms each variable into a uniform distribution. This type of the redundancy/mutual information estimate, even its coarse-grained version, is invariant against any monotonous (possibly nonlinear) transformation of data [138] and in consequence, information-theoretic functionals estimated by this algorithm are independent of the univariate distribution of the input data.

Now, let the n variables $X_1, \ldots, X_n$ have zero means, unit variances and the correlation matrix $\mathbf{C}$. Then, we define the *linear redundancy* $L(X_1; \ldots; X_n)$ of $X_1, X_2, \ldots, X_n$ as

$$L(X_1; \ldots; X_n) = -\frac{1}{2} \sum_{i=1}^{n} \log(\sigma_i), \quad (18)$$

where σ_i are the eigenvalues of the $n \times n$ correlation matrix $\mathbf{C}$.

If $X_1, \ldots, X_n$ have an n-dimensional Gaussian distribution, then $L(X_1; \ldots; X_n)$ and $R(X_1; \ldots; X_n)$ are theoretically equivalent.

In practical applications one deals with a time series $\{y(t)\}$, considered as a realization of a stochastic process $\{Y(t)\}$, which is stationary and ergodic. Then, due to ergodicity, all the subsequent information-theoretic functionals are estimated using time averages instead of ensemble averages, and the variables X_i are substituted as

$$X_i = y(t + (i-1)\tau). \quad (19)$$

Due to stationarity the redundancies

$$R^n(\tau) \equiv R(y(t); y(t+\tau); \ldots; y(t+(n-1)\tau)) \quad (20)$$

and

$$L^n(\tau) \equiv L(y(t); y(t+\tau); \ldots; y(t+(n-1)\tau)) \quad (21)$$

are functions of n and τ, independent of t.

Let us remind also equivalent definitions of the redundancy, or, for simplicity of the mutual information of two discrete random variables X and Y with sets of values Ξ and Υ, respectively, and probability distribution functions $p(x)$, $p(y)$ and joint PDF $p(x, y)$. The *entropy* $H(X)$ of a single variable, say X, is defined as

$$H(X) = -\sum_{x \in \Xi} p(x) \log p(x), \quad (22)$$

and the *joint entropy* $H(X,Y)$ of X and Y is

$$H(X,Y) = -\sum_{x\in\Xi}\sum_{y\in\Upsilon} p(x,y)\log p(x,y). \quad (23)$$

The *conditional entropy* $H(Y|X)$ of Y given X is

$$H(Y|X) = -\sum_{x\in\Xi}\sum_{y\in\Upsilon} p(x,y)\log p(y|x). \quad (24)$$

The average amount of common information, contained in the variables X and Y, is quantified by the *mutual information* $I(X;Y)$, defined as

$$I(X;Y) = H(X) + H(Y) - H(X,Y). \quad (25)$$

The mutual information normalized as

$$\iota(X;Y) = \frac{I(X;Y)}{\max(H(X),H(Y))} \quad (26)$$

attains values between 0 and 1, and can be used to define a distance measure $d(X,Y)$ as

$$d(X,Y) = 1 - \iota(X;Y), \quad (27)$$

which has all mathematical properties of a distance in the space of random variables [139]. Thus $d(.,.)$ define a metric based on the strength of dependence. Independent variables have the maximum distance ($d(.,.) = 1$), the functionally related variables have a zero distance.

The conditional mutual information $I(X;Y|Z)$ of the variables X, Y given the variable Z is given as

$$I(X;Y|Z) = H(X|Z) + H(Y|Z) - H(X,Y|Z). \quad (28)$$

For Z independent of X and Y we have

$$I(X;Y|Z) = I(X;Y). \quad (29)$$

By a simple manipulation we obtain

$$I(X;Y|Z) = I(X;Y;Z) - I(X;Z) - I(Y;Z). \quad (30)$$

Thus the conditional mutual information $I(X;Y|Z)$ characterizes the "net" dependence between X and Y without the possible influence of another variable, Z.

The entropy and information are usually measured in bits if the base of the logarithms in their definitions is 2, here we use the natural logarithm and therefore the units are called nats.

4.4. The null hypothesis of nonlinearity tests and its negations

We have already mentioned some problems embedded in the nonlinearity tests based on the FT surrogate data. We should, however, return to the basic null hypothesis and precisely consider its negations before interpreting possible rejections of the null hypothesis. The null hypothesis realized by the FT surrogates is in principle equivalent to a linear stochastic process such as that described by the ARMA model (10). It is very common in the nonlinear dynamics literature to consider the rejection of the null (10) as an evidence for a process such as (13) with a nonlinear skeleton (11). This is, however, only one of possible negations of (10). A number of different processes should be considered, which possess a linear deterministic skeleton[232], i.e., a linear AR part – the first sum in (10), or no deterministic skeleton at all (MA processes), however, their innovations $\{N(t)\}$ do not fulfill the conditions given above. Generally, one or more of the following properties could reject the null (10):

1. The innovations $\{N(t)\}$ are not Gaussian.
2. The innovations $\{N(t)\}$ are not an iid process, where iid means that the innovations should be not only uncorrelated, but generally independent.
3. The variance of $\{N(t)\}$ is not constant.

We will demonstrate such effects in the following examples.

5. TESTING FOR NONLINEARITY IN THE ATMOSPHERIC DYNAMICS

Complex dynamical phenomena in the atmosphere and the prediction of the weather or the climate have been challenging physicists for decades. It is not surprising that many papers have been published, devoted to the problem of inferring the dynamical mechanisms of the weather and of the climate changes from recorded data using the ideas and methods of the theory of deterministic chaos. The measured quantities, selected for the analyses, have included, e.g., local surface pressures, relative sunshine durations, zonal wave amplitudes [140], upper-level geopotential heights [141, 142], low-level vertical velocity components [143], or, oxygen-isotope concentrations in deep sea cores [140, 144–147]. In the majority of the cases the Grassberger-Procaccia algorithm for estimating the correlation dimension [51] was used as the analysis tool, and obtained low values of the dimension estimates were claimed as evidence for low-dimensional chaos in the weather or climate dynamics [140–144, 146]. On the other hand, Grassberger [145] cautioned, that in the case of short and noisy data, as the climatic and weather records usually are, the reliability of the method is questionable and the low values of the dimension estimates may be spurious. And indeed, he constructed a random series of corresponding length, preprocessed by the same way as the climatic record in [144] and obtained a low value of the estimated dimension. Also Lorenz [148] writes that it seems unlikely that the global weather or climate systems possess a low-dimensional attractor. As a contribution to this discussion, Paluš and Novotná [149] applied a test for nonlinearity to weather related data.

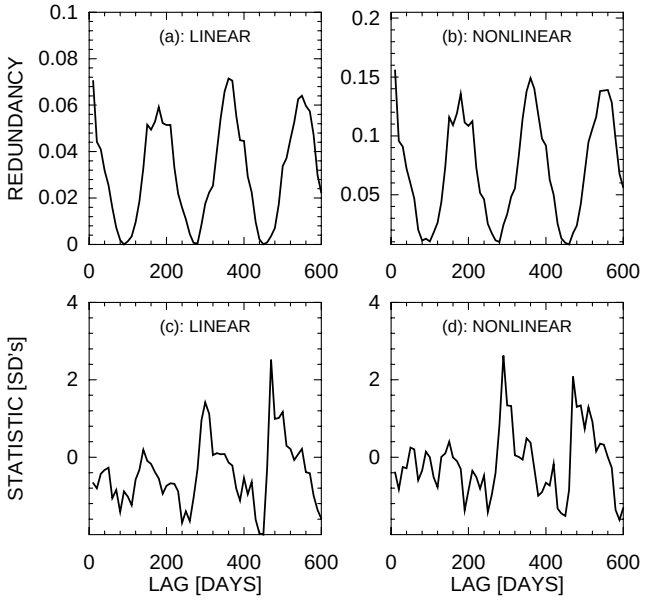

FIG. 8: (a) Linear redundancy $L^2(\tau)$, (b) redundancy $R^2(\tau)$, as functions of the lag τ, (c) linear redundancy statistic, (d) redundancy statistic for the Prague series of the geopotential heights of 500 hPa isobaric level. Embedding dimension $n = 2$.

In their test the above information-theoretic functionals were used as the discriminating statistic with the FT surrogate data.

5.1. Lack of nonlinearity in geopotential heights and in the temperature

In Ref. [149], two series of daily values of geopotential heights of 500 hPa isobaric level were analysed, the first, 6570 samples (18 years) recorded in Prague, Ruzyně station; the second, 11670 samples (32 years) recorded in Krakow. The surrogate data were generated using the fast Fourier transform (FFT), which requires the series length to be a power of two, therefore subseries of lengths 4096 and 8192 samples, respectively, were analysed.

The other two series, analysed in [149], were recorded in Prague, Klementinum station, and are more unique: the series of 200 years (73000 samples) of mean daily values of the near-surface air temperature and daily values of the near-surface atmospheric pressure. Again, due to the FFT-based surrogates, the subseries of 65536 samples were analysed.

Paluš et al. [150, 151] demonstrated that already a qualitative comparison of the graphs of the redundancy $R^n(\tau)$ and the linear redundancy $L^n(\tau)$, as functions of the time lag τ, can serve as an indication of nonlinearity. This comparison, referred to as the *qualitative testing*, is then followed by the standard surrogate data test in which the redundancy $R^n(\tau)$ serves as the discriminating statistic. The same test, as that using $R^n(\tau)$, is done using the linear redundancy $L^n(\tau)$. Since $L^n(\tau)$ is given by the autocorrelation function – see Eq. (18), the application of $L^n(\tau)$ as a discriminating statistic tests the quality of the surrogate data. A non-significant test with $L^n(\tau)$ confirms that the autocorrelation function of the tested data is preserved in the surrogates within the statistical accuracy. A significance in this test identifies a problem in the surrogate data, so that a possible significance obtained using a nonlinear statistic can be spurious.

The results of the analysis of the Prague series of the geopotential heights are presented in Fig. 8. The qualitative comparison shows no substantial difference between the time plots of the linear redundancy $L^2(\tau)$ and the redundancy $R^2(\tau)$, showed in Figs. 8a and 8b, respectively. In the quantitative analysis with the FT surrogates (Figs. 8c, d), where the discriminating statistic is defined as the difference between $R^2(\tau)$ (or $L^2(\tau)$) from the data and the mean values from the surrogate set, given in the number of the surrogate SD's, there are several formally significant results (i.e. differences greater than 1.699SD), however, even if we suppose a normal distribution of this statistic, there are two reasons why not to reject the null hypothesis of a linear stochastic process:
a) Statistical reason: Due to multiplicity of the test values (60 in this case) the criterion for significance of an individual value must be strengthened, i.e., based on the Bonferroni inequality we should take, in this case, $\alpha < 0.05/60$ instead of $\alpha < 0.05$ [152, 153], what increases the critical value of the statistics from 1.699 to approximately 3.5. This approach, however, is fully correct for independent test values, for the dependent test values, what is the case here, the power of the test can be decreased. In order to avoid the type II error (i.e., acceptance of the null hypothesis when it should be rejected), the Heilperin-Ruger inequality can be considered instead of the Bonferroni inequality, and, expecting k significant values (from m total test values) $\alpha < 0.05k/m$ can be taken [154, 155]. In this case the critical value is still about 3. Thus, no significant difference was found.
b) Methodological reason: Even if we accept some values of the nonlinear statistic (based on $R^n(\tau)$) as significant, we cannot reliably reject the null hypothesis of a linear stochastic process, as far as equivalent differences were found in the linear statistic (based on $L^n(\tau)$). Therefore the observed differences can be caused by the fact that the surrogates do not exactly mimic the linear properties of the data, and not necessarily by a nonlinearity. Therefore, using both the qualitative and quantitative methods, Paluš and Novotná [149] concluded that the Prague series of the geopotential heights was not unambiguously discernible from the isospectral linear stochastic process.

The main feature of the dynamics of the above geopotential heights series, as it can be observed in the time-lag plots of the redundancies (Fig. 8a, b), is the one-year periodicity. Asking whether there is anything beyond this dynamics, also the filtered series was analysed, in which one-year periodicity was eliminated by the FFT based filter. In the qualitative analysis both $L^n(\tau)$ and

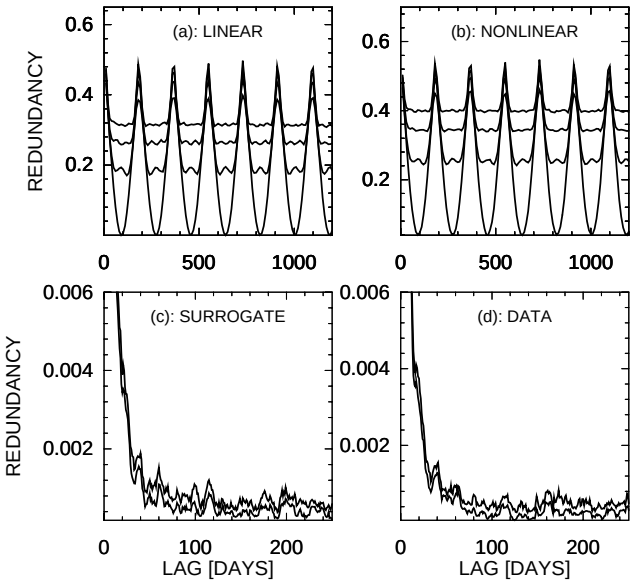

FIG. 9: (a) Linear redundancy $L^n(\tau)$, (b) redundancy $R^n(\tau)$, computed from the near-surface air temperature series. The four different curves in each picture are the redundancies for different embedding dimensions, $n = 2 - 5$, reading from the bottom to the top. Redundancies $L^n(\tau)$ and $R^n(\tau)$ are plotted as $L^n(\tau)/(n-1)$ and $R^n(\tau)/(n-1)$, respectively. (c) $R^n(\tau)$, $n = 2$ (lower curve), 3 (upper curve), for the surrogate data of the filtered temperature series, (d) $R^n(\tau)$, $n = 2$ (lower curve), 3 (upper curve), for the filtered temperature series.

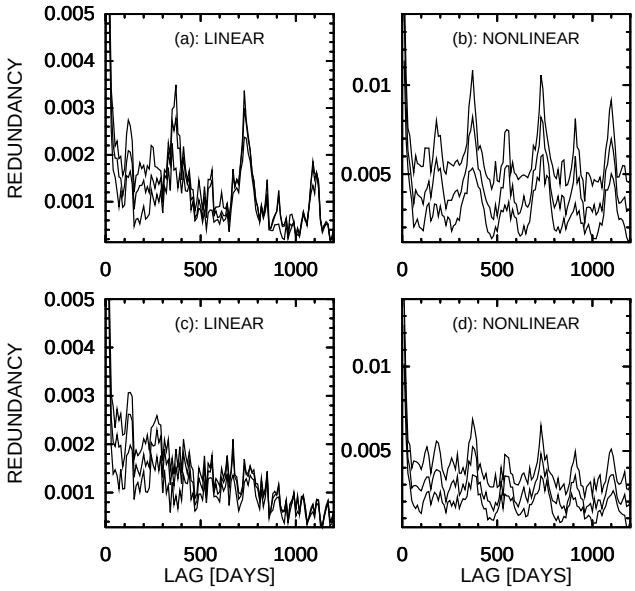

FIG. 10: (a) Linear redundancy $L^n(\tau)$, (b) redundancy $R^n(\tau)$, computed from the near-surface air pressure series. (c) Linear redundancy $L^n(\tau)$, (d) redundancy $R^n(\tau)$, computed from the filtered near-surface air pressure series. The three different curves in each picture are the redundancies for different embedding dimensions, $n = 2 - 4$, reading from the bottom to the top. Redundancies $L^n(\tau)$ and $R^n(\tau)$ are plotted as $L^n(\tau)/(n-1)$ and $R^n(\tau)/(n-1)$, respectively.

$R^n(\tau)$ show the same picture – they decrease quickly until the lag 12 days and then fluctuate about a very low level. The question whether these small values (2 – 4 $\times 10^{-3}$) can mean a "numerical zero", i.e., the fact that the filtered series $\{y(t)\}$ and $\{y(t+\tau)\}$ for $\tau > 12$ are independent, was answered by the quantitative test using the scrambled surrogates – the elements of the series were mixed in temporal order so that all temporal correlations were destroyed. Comparing the data with the scrambled surrogates the null hypothesis of an IID process was tested an rejected (differences of 4 – 8 SD's). On the other hand, using the FFT surrogates, both the stronger dependences (the lags 1 – 12) and the weak dependences for the lags $\tau > 12$ days were found consistent with the isospectral linear stochastic process.

The results for the Krakow series were very similar to those for the above Prague series. Therefore, Paluš and Novotná [149] concluded that the analysis of the recordings of the geopotential heights did not yield sufficient arguments for rejecting the linear stochastic explanation.

The results of the analysis of the surface air temperature record (65,536 samples – days) are presented in Fig. 9. The qualitative analysis of the data (Figs. 9a, b) brought no substantial difference between $R^n(\tau)$ and $L^n(\tau)$, $n = 2 - 5$, $\tau = 10 - 1200$ days (Figs. 9a, b). In the quantitative analysis, the differences obtained were not greater than 1.6 SD's. After filtering out the one-year periodicity, the quantitative analysis brought no significant results, like the analysis before the filtration. In Figs. 9c, d we can see that the redundancies of the filtered temperature series decrease until the lag of about 80 days and then fluctuate about the same (low) level. Similarly, like in the case of the geopotential heights data, the hypothesis of an IID process was rejected, however, the observed dependence was found consistent with a linear stochastic explanation. The analysis of the temperature record brought no arguments to reject the null hypothesis of a linear stochastic process. Note that in this case the linear redundancy $L^n(\tau)$ did not detect any significant differences from the surrogate data, although the temperature data contains a strong periodic component. As discussed by Paluš [132], the above-mentioned problems with reproducing the autocorrelation function of the cyclic data emerges in relatively short time series such as the series of the geopotential heights, while using long time series (such as the temperature series of 65536 samples) the FFT algorithm can produce reliable surrogate data.

5.2. Pressure data: What kind of nonlinearity do we observe?

The results of the analysis of the near-surface air pressure record (65,536 samples – days) are presented in Fig. 10. The qualitative analysis of the data (Figs. 10a, b) shows some differences between $L^n(\tau)$ and $R^n(\tau)$,

namely the half-year peaks are not so clearly pronounced in $L^n(\tau)$ than in $R^n(\tau)$, the one-year periodicity, however, is apparent in both $L^n(\tau)$ and $R^n(\tau)$. More clear results were obtained by the quantitative analysis: While no significant differences were detected on the linear level, i.e., by the statistic based on $L^n(\tau)$, the nonlinear statistic (based on $R^n(\tau)$) brought significant differences of values between 5 and 15 SD's. The results of the analysis of the filtered pressure series are even more illustrative: The results of the quantitative analysis did not change after the filtration, i.e., evidence for nonlinearity, safe from spurious effects of differences on the linear level, was detected. In the qualitative comparison (Figs. 10c, d), we can see that $R^n(\tau)$ decreased after the filtration, i.e., the linear contribution to the dependence structures in the data (reflected in nonlinear $R^n(\tau)$ as well) was removed by the filtration, while the character of the time-lag dependence of $R^n(\tau)$ is almost the same as in $R^n(\tau)$ computed from the original data, i.e., the principal one-year peaks and the smaller half-year peaks were detected. On the other hand, linear redundancy $L^n(\tau)$ of the filtered data does not reveal these structures. Note that the filtering in Ref. [149] was done in the spectral domain. We can analyse the pressure data again using so-called daily anomalies, defined as the differences from the long-term daily averages. This transformation of data (almost entirely) removes the annual cycle as the seasonality in the mean and does not cause effects sometimes observed after using the filters in the spectral domain. The results are presented in Fig. 11. For lags larger than a few days there is only a weak linear dependence, as measured by the linear redundancy (Fig. 11a) and reflected in the surrogates, but the (nonlinear) redundancy detects a clear dependence as an oscillatory structure with a yearly periodicity. This difference is highly significant (10 – 30 SD's, Fig. 11d), while no significance in the linear statistic (Fig. 11c) confirms the quality of the surrogates, which correctly reflect 'the linear properties' of the data. So we confirm the nonlinearity detected in the previous analysis (Fig. 10). Can this result be understood as an evidence for the model (13) with a nonlinear periodic skeleton F, which could provide predictability of the air pressure for several years in advance?

The seasonality in the mean present in this data (Fig. 12, upper panel) was mostly removed by considering the differences from the long term daily averages. The problem is that also the variance of this data is not constant, but clearly seasonal (Fig. 12, the lower panel, the standard deviation (SD) is the square root of the variance). This property is 'nonlinear' in the sense that the surrogate data and the model (10) possess a constant variance and cannot reproduce the seasonality in the variance. After rescaling the data in order to obtain a constant variance, the effect of the false long-term nonlinear dependence is lost (Fig. 13). In the rescaled data there is no long-term dependence except of a weak linear link due to the not entirely removed seasonality in the mean, i.e., although the null hypothesis (10) was rejected,

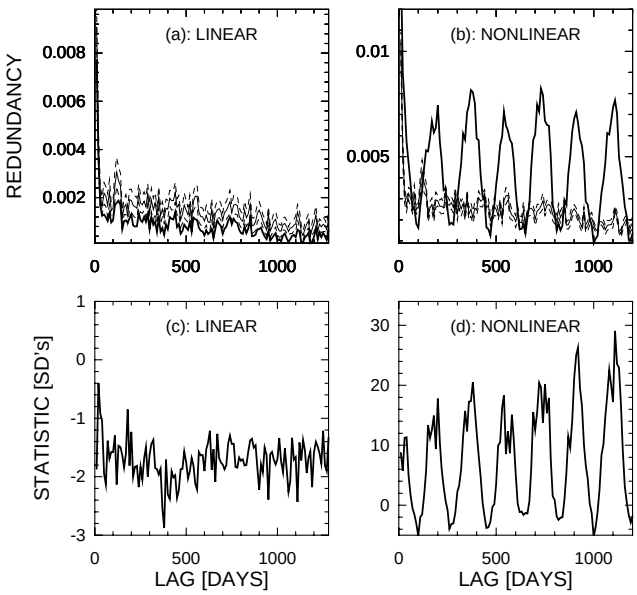

FIG. 11: (a) Linear redundancy $L(y(t); y(t+\tau))$ (solid line), (b) redundancy $R(y(t); y(t+\tau))$ (solid line), for the series of the differences from the long term daily averages of the near-surface air pressure (Prague-Klementinum station) and for its FT surrogates (thin solid and dashed lines present mean and mean±SD, respectively, of a set of 30 surrogate realizations); (c) linear (L-based), and (d) nonlinear (R-based) statistics; as functions of the time lag τ, measured in days.

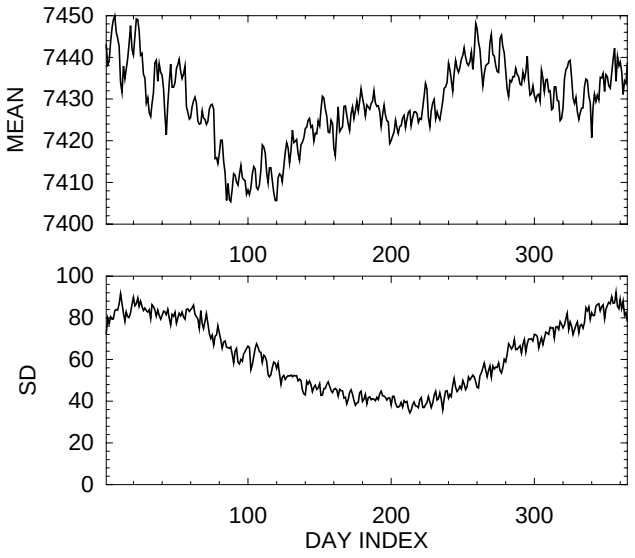

FIG. 12: Seasonality in the mean (upper panel) and in the variance (lower panel) of the near-surface air pressure series – long term means (upper panel) and standard deviations (square root of variance, lower panel) for each day in a year. The days are consecutively numbered, January 1 has the index 1, January 2 has the index 2, ..., February 1 has the index 32, etc.

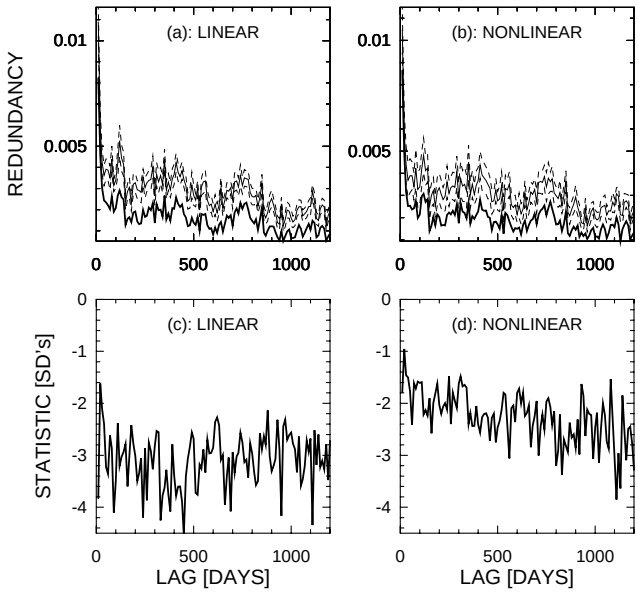

FIG. 13: (a) Linear redundancy $L(y(t); y(t+\tau))$ (solid line), (b) redundancy $R(y(t); y(t+\tau))$ (solid line), for a series of differences from the long term daily averages of the near-surface air pressure, rescaled in order to have a constant variance, and for its FT surrogates (thin solid and dashed lines present mean and mean±SD, respectively, of a set of 30 surrogate realizations); (c) linear (L-based), and (d) nonlinear (R-based) statistics; as functions of the time lag τ, measured in days.

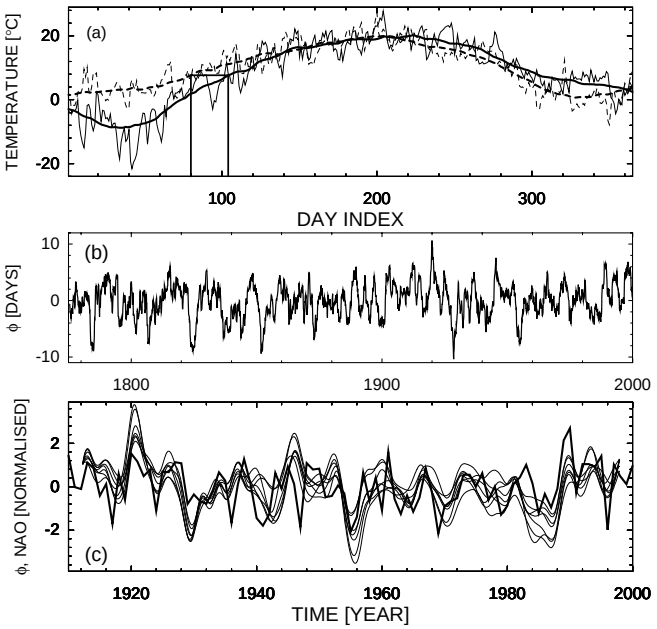

FIG. 14: (a) The daily mean near-surface air temperature in Prague in 1920 (dashed lines) and in 1929 (solid lines), raw data (thin lines) and smoothed data (thick lines). The solid straight lines show that the temperature of March 21, 1920 was reached with a delay of 24 days in 1929. (b) The phase $\phi(t)$ of the annual temperature cycle obtained from the Prague record (1775 – 2001). (c) The phase $\phi(t)$ of the annual temperature cycle obtained from seven European stations (thin smooth lines); and the North Atlantic Oscillation index averaged over the January-March period (thick line). All data are normalized to zero mean and unit variance.

the rejection was apparently caused by the seasonal (non-constant) variance and not by a nonlinear deterministic structure (11).

The above example of the near-surface air pressure clearly demonstrated the influence of variable variance on the redundancy – surrogate data nonlinearity test [132]. The effect of non-Gaussian innovations $\{N(t)\}$ was discussed in [132], and a possible influence of non-iid $\{N(t)\}$ (i.e., innovations containing (nonlinear) temporal structures) is understandable. It is important to note that similar effects of 'defective' innovations in a process under study would effect not only this particular test for nonlinearity, but all tests which use some type of FT/ARMA surrogates, and also any method which contain the process (10) at least implicitly in its construction. Also, all entropy-related statistics, that is, not only the above information-theoretic functionals, but also, for instance, statistics based on the correlation integral [156], are extremely sensitive to variable variance and/or to (non)Gaussianity of the innovations $\{N(t)\}$. Therefore one should very carefully assess results of nonlinearity tests in order to avoid confusing this kind of effects with actual nonlinear functional dependence in the data under study. Some authors [157] even report that the 'strength of nonlinearity' in air pressure data depends on latitude, with the weakest or no nonlinearity in tropical or subtropical regions. The influence of the seasonality in the variance should be carefully assessed before taking such conclusions for granted.

5.3. But the atmosphere is not linear ...

We should again reiterate the statement that the lack of arguments for rejecting the null hypothesis is not an evidence that the null hypothesis is true. The above results or the example of Grassberger [145] do not say that the dynamics of the atmosphere is linear, they just disqualify the assertion that a low-dimensional, nonlinear, chaotic dynamical system in the evolution of the weather and climate was detected. Neither the correlation dimension, nor the mutual information unambiguously detect a functional nonlinearity in the analysed meteorological data. The statistical tests with the surrogate data prevent researchers from making claims not supported by the available data. Even this 'negative' result is useful in the process of broadening scientific knowledge. Can we, however, use the testing with the surrogate data in obtaining new knowledge? Of course, we can, we just need to specify more focused questions. In the next Section 6 we will demonstrate that it is more useful to test for a specific nonlinear property, than just to test for general, unspecified nonlinearity.

Practically all atmospheric data are influenced by the annual cycle. The latter is most prominent in air temperature series recorded in mid and higher latitudes. Let us

consider that the dominant component of the annual temperature cycle can be written, apart from a noise term, as

$$T(t) = A(t)\cos[\omega t + \phi(t)], \tag{31}$$

where t is time, $A(t)$ is the amplitude, ω is the (constant) frequency given by the tropical year. The phase $\phi(t)$ describes the difference from the exact annual cycle, so that it reflects fluctuations in the timing of seasons. Paluš et al. [158] demonstrated that the phase fluctuations $\phi(t)$ are rather large, giving differences in starting of the spring seasons up to 24 days inside a decade (Fig. 14). Thus, on the century scale, a phase shift due to the precession is negligible in comparison with the observed phase fluctuations $\phi(t)$.

If the atmospheric dynamics was linear stochastic, the phase fluctuations $\phi(t)$ could be well described by a random process. Paluš et al. [158] show that these fluctuations are not arbitrary, but are correlated with a large-scale atmospheric process, known as the North Atlantic Oscillation [159]. The inherent dynamics of the phase fluctuations $\phi(t)$ was not analysed yet, however, it is a challenging task, since it could demonstrate a nonlinear response of the atmosphere to the regular driving given by the movement of the Earth. Uncovering a dynamical mechanism underlying the phase fluctuations $\phi(t)$ would also allow predictions of the fluctuations of seasons. Potential skills in the prediction of the onsets of seasons could have significant socio-economic impacts, while an unpredictable phase in the climate may be a more serious problem to society than changes in the amplitude of the annual cycle or even of the mean temperature [160], considering the strong influence of seasons at mid and higher latitudes on natural ecosystems and human activities such as agriculture, forestry, water management, transportation and tourism.

6. NONLINEARITY IN THE SUNSPOT NUMBERS

Let us return to the famous sunspot numbers, probably the longest historical record of the solar variability. As we mentioned in Sec. 3 3.3, in 1852 Wolf [115] reported the now well-know 11-year cycle. Of course, the sunspot cycle is not strictly periodic, but fluctuations in its amplitude as well as in its frequency (i.e., in the cycle duration) occur. Regularities and irregularities in the solar activity cycle [161] are among the most intriguing and poorly understood aspects of the Sun. Dynamo theory [162–164], describing complex magnetohydrodynamic plasma motions inside the Sun, has resulted in many models which reproduce basic features of the solar activity [165–167]. However, the nature of the solar cycle is far from being understood and the dynamo models are not predictive [168]. In many cases stochastic models were constructed in order to make predictions of a future behaviour of the sunspot cycle (see [169] and references therein). On the other hand, development in nonlinear dynamics and the theory of deterministic chaos, namely the methods and algorithms for analysis and prediction of nonlinear and chaotic time series have naturally found their way into the analyses of the sunspot series. Several authors [38–40] have claimed an evidence for a deterministic chaotic origin of the sunspot cycle, based on estimations of correlation dimension, Lyapunov exponents and an increase of a prediction error with a prediction horizon. However, as we discussed above (Sec. 2 2.2), the dimensional algorithms have been found unreliable when applied to relatively short experimental data, and properties consistent with stochastic processes (colored noises) such as autocorrelations can lead to a spurious convergence of dimensional estimates [70, 73, 78, 79]. Similar behavior has been observed also for Lyapunov exponent estimators [170, 171]. And the increase of a prediction error with an increasing prediction horizon is not a property exclusive for chaos, but it can also be observed in stochastic systems. Therefore, such results cannot be considered as a convincing evidence for a nonlinear dynamical origin of the sunspot cycle. For instance, Theiler et al. [83] demonstrate that the correlation dimension cannot distinguish the sunspot data from their AAFT surrogate data, while a prediction error statistic gives a weak distinction (2–3SD). Only a simple skew statistic, measuring the temporal asymmetry of a signal, is able to reliably reject the AAFT surrogate data [83]. The asymmetric behaviour of the sunspot cycle, however, can be mimicked by a transformation of a linear stochastic model, described below (the Barnes model).

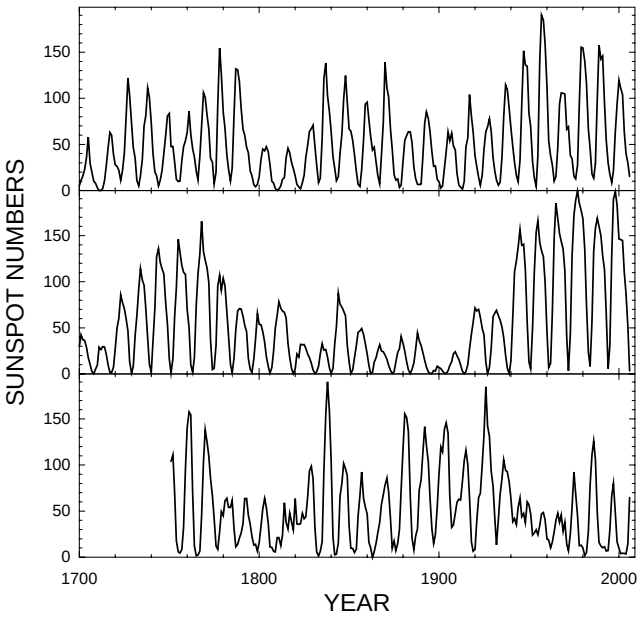

FIG. 15: The yearly sunspot numbers 1700–2006 (top panel). A 307-sample realization of the Barnes model (middle panel). A realization of the IAFT surrogate data of a 256-sample subset of the sunspot numbers (bottom panel).

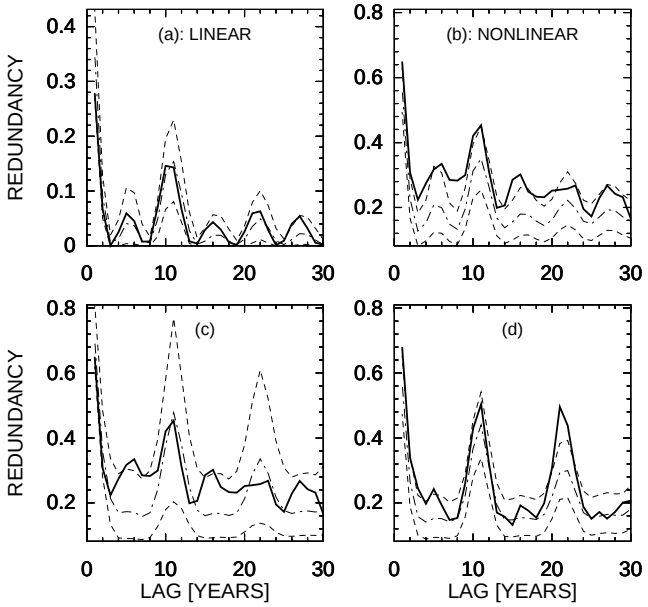

FIG. 16: (a) Linear redundancy $L(y(t); y(t+\tau))$ (solid line) for the series of sunspot numbers as a function of the lag τ; linear redundancy $L(y(t); y(t+\tau))$ for the related IAFT surrogate set: dash-and-dotted line presents the mean, dashed lines illustrate the 2.5th and the 97.5th (bottom and top line, respectively) percentile of the surrogate $L(y(t); y(t+\tau))$ distribution. (b) Nonlinear (general) redundancy $R(y(t); y(t+\tau))$ (solid line), for the series of sunspot numbers as a function of the lag τ; nonlinear redundancy $R(y(t); y(t+\tau))$ for the related IAFT surrogate set, the same line codes as in (a). (c) Nonlinear redundancy $R(y(t); y(t+\tau))$ (solid line), for the series of the sunspot numbers as a function of the lag τ; nonlinear redundancy $R(y(t); y(t+\tau))$ for the Barnes model surrogate set, the same line codes as in (a). (d) Nonlinear redundancy $R(y(t); y(t+\tau))$ (solid line), for a realization of the Barnes model as a function of the lag τ; nonlinear redundancy $R(y(t); y(t+\tau))$ for the related IAFT surrogate set, the same line codes as in (a).

Let us demonstrate the test for nonlinearity using the information-theoretic functionals as discriminating statistics in the case of the series of the yearly sunspot numbers from the period 1700 – 2006 (Fig. 15, top panel). At the first step we use the iterative, amplitude adjusted FT (IAFT) surrogate data (Fig. 15, bottom panel).

As the test results, here we do not present differences from surrogate mean in number of standard deviations as in the previous section, but using 15,000 surrogate replications we estimate the 2.5th and 97.5th percentiles of distributions of the redundancies computed from the surrogate data. If the values of the redundancy obtained from the studied data lies outside the range given by these two percentile values, the null hypothesis can be rejected on the significance level $\alpha < 0.05$.

The linear redundancy $L(y(t); y(t+\tau))$ (solid line in Fig. 16a) lies clearly inside the 2.5th and the 97.5th percentiles of the $L(y(t); y(t+\tau))$ IAFT surrogate distribution (dashed lines in Fig. 16a). This test is just the check of the quality of the surrogate data and it says that the linear properties (dependence structures) in the sunspot data do not differ from those of the IAFT surrogates (a realization of which is presented in Fig. 15, bottom panel), so that the surrogates should not be the source of a spurious detection of nonlinearity.

The nonlinearity test itself is presented in Fig. 16b, where the (nonlinear) redundancy $R(y(t); y(t+\tau))$ (solid line in Fig. 16b) is, for majority of the studied lags, higher than the 97.5th percentile of the IAFT surrogate distribution (the upper dashed line in Fig. 16b). Thus the null hypothesis of a linear stochastic process (10), possibly passed through a static nonlinear transformation, is rejected.

Does this rejection really mean that a nonlinear dynamical system such as (11) or (13) underlies the sunspot cycle, or can this rejection be explained by any of the reasons listed in Sec. 4 4.4 ? This question is hard to answer. For instance, we cannot evaluate properties of innovations (model residuals) without an a-priory knowledge of a valid model. Physical models trying to explain the variation of the solar activity come from the dynamo theory [164]. The principle of such a self-exciting dynamo is that the magnetic field is amplified and maintained by the interaction of mainly three types of hydrodynamic plasma motions, namely differential rotation, turbulent convection and helicity. It is interesting to mention that there are some rather simple conceptual dynamo models which show a rich dynamical behaviour and can explain several facts known from observations [165]. Such models, however, are not fitted directly to experimental data, are evaluated only in a qualitative way, and are not predictive.

The Barnes model [126]:

$$z_i = \alpha_1 z_{i-1} + \alpha_2 z_{i-2} + a_i - \beta_1 a_{i-1} - \beta_2 a_{i-2}, \qquad (32)$$

$$s_i = z_i^2 + \gamma (z_i^2 - z_{i-1}^2)^2, \qquad (33)$$

where $\alpha_1 = 1.90693$, $\alpha_2 = -0.98751$, $\beta_1 = 0.78512$, $\beta_2 = -0.40662$, $\gamma = 0.03$ and a_i are IID Gaussian random variables with zero mean and standard deviation SD=0.4; is a simple but efficient model able to mimic essential statistical properties of the sunspot numbers. It incorporates the structure of an autoregressive moving average ARMA(2,2) model (32) with a nonlinear transformation (33) which ensures that the generated series remains asymmetric and positive and tends to increase more rapidly than it decreases. Moreover, the stochastic Barnes model can mimic some seemingly nonlinear properties such as behaviour of correlation integrals [173] and phase portraits [174] obtained from the sunspot series.

We can evaluate, in the sense of the above nonlinearity test, whether the Barnes model can explain the sunspot data, by using realizations of the Barnes model as the surrogate data (Fig. 15, middle panel). The result of such a test is presented in Fig. 16c.

The nonlinear redundancy $R(y(t); y(t+\tau))$ (solid line in Fig. 16c) only slightly exceeds the 97.5th percentile of

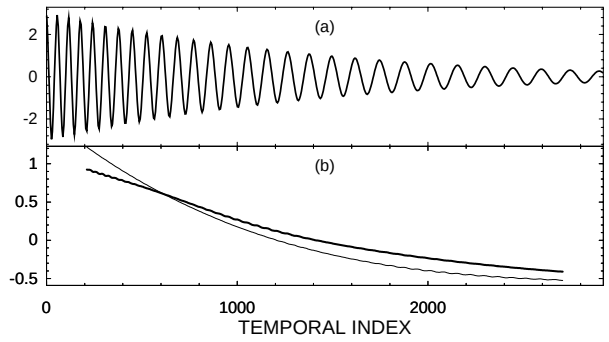

FIG. 17: (a) A solution of the nonlinear Duffing oscillator without any external driving force, and (b) the related instantaneous amplitude (solid line) and frequency (thin line).

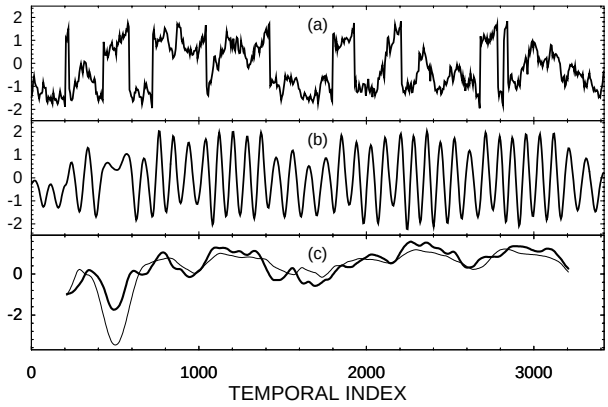

FIG. 18: (a) A random driving force (a random walk with a few jumps). (b) A solution of the nonlinear Duffing oscillator with the random driving force $F(t)$ plotted in panel (a); (c) instantaneous amplitude (solid line) and frequency (thin line) extracted from the solution in panel (b).

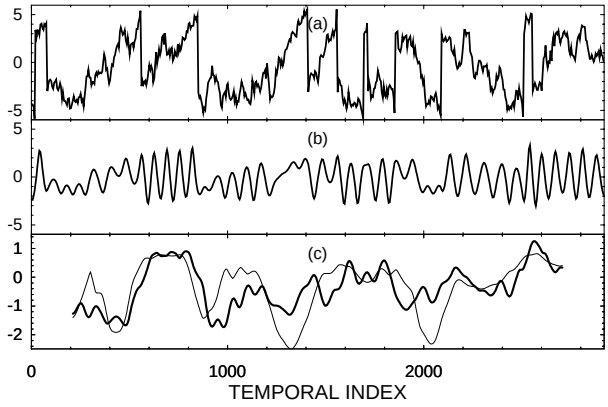

FIG. 19: (a) Another example of a random driving force ("stronger", i.e., with higher amplitude than in the previous case in Fig. 18). (b) A solution of the nonlinear Duffing oscillator with the random driving force $F(t)$ plotted in panel (a); (c) instantaneous amplitude (solid line) and frequency (thin line) extracted from the solution in panel (b).

the Barnes surrogate distribution in 3 from the 30 studied lags. Considering the problem of simultaneous statistical inference, discussed above (Sec. 5 5.1), this result is not sufficient for the rejection of the null hypothesis presented by the Barnes model. If we do not reject the Barnes model using the nonlinear redundancy $R(y(t); y(t+\tau))$, the linear redundancy $L(y(t); y(t+\tau))$ would not reject it either, since it is a special case of $R(y(t); y(t+\tau))$. This means that the linear properties (the autocorrelation function) of realizations of the Barnes model are consistent with the linear properties of the original data.

On the other hand, when we test for nonlinearity in a realization of the Barnes model (Fig. 16d), the IAFT surrogates are rejected. The rejection is clear only in the lags 21 and 22, however, $R(y(t); y(t+\tau))$ of the tested series there exceeds the whole range of the surrogate values, i.e., the significance $p = 0$ and the test is significant even considering the problem of the simultaneous statistical inference. This result could be expected, since the nonlinear transformation (33) is not static, therefore the IAFT surrogates are rejected.

To summarize the last two tests, the realization of the Barnes model appeared in the nonlinearity test as nonlinear, and the Barnes model as the null hypothesis for an explanation of the dependence structures in the sunspot cycle cannot be unambiguously rejected. Can we find any solid argument for a nonlinear dynamical origin of the sunspot cycle, or should we accept a linear stochastic explanation, such as the Barnes model?

6.1. Amplitude-frequency correlation in nonlinear oscillators

In this section we demonstrate a typical property of nonlinear oscillators (a class of nonlinear dynamical systems), namely the correlation between the instantaneous amplitude and the instantaneous frequency of signals generated as solutions of such systems.

As a demonstrative example of a nonlinear oscillator (not a model for the sunspot cycle) we will consider the Duffing oscillator

$$\ddot{x} + 0.05\dot{x} + x + x^3 = F(t). \tag{34}$$

If $F(t) = 0$ and without the cubic member x^3, the equation (34) represents a damped linear oscillator with a constant frequency and an exponentially decreasing amplitude. The presence of the nonlinear (cubic) member x^3 in the equation (34) leads to a time dependent frequency, and considering again $F(t) = 0$, both the amplitude $A(t)$ and frequency $\omega(t)$ exponentially decrease and are correlated (Figs. 17a,b).

Now, consider that the nonlinear oscillator (34) is driven by a random driving force $F(t)$. In the numerical examples presented here we consider a simple random walk with a few jumps as the driving force $F(t)$.

The relation between $A(t)$ and $\omega(t)$ is a nonlinear function and may vary in time, however, the level of

the correlation between $A(t)$ and $\omega(t)$ depends on the driving force: With a relatively weak driving (Fig. 18a), $A(t)$ and $\omega(t)$ are almost perfectly correlated (Fig. 18c), with a stronger driving force $F(t)$ (Fig. 19a) some differences between $A(t)$ and $\omega(t)$ emerge, however, $A(t)$ and $\omega(t)$ are still correlated (Fig. 19c).

The above presented amplitude-frequency correlation is a property which can be tested in experimental signals, even in scalar cases (single measured time series). The instantaneous amplitude and phase of a signal $s(t)$ can be determined by using the analytic signal concept of Gabor [175], introduced into the field of nonlinear dynamics within the context of phase synchronization by Rosenblum et al. [176]. The analytic signal $\psi(t)$ is a complex function of time defined as

$$\psi(t) = s(t) + j\hat{s}(t) = A(t)e^{j\phi(t)}, \tag{35}$$

where the function $\hat{s}(t)$ is the Hilbert transform of $s(t)$

$$\hat{s}(t) = \frac{1}{\pi}\ \text{P.V.}\int_{-\infty}^{\infty}\frac{s(\tau)}{t-\tau}d\tau. \tag{36}$$

(P.V. means that the integral is taken in the sense of the Cauchy principal value.) $A(t)$ is the instantaneous amplitude and the instantaneous phase $\phi(t)$ of the signal $s(t)$ is

$$\phi(t) = \arctan\frac{\hat{s}(t)}{s(t)}. \tag{37}$$

The instantaneous frequency $\omega(t)$ is the temporal derivative $\dot{\phi}(t)$ of the instantaneous phase $\phi(t)$.

6.2. Amplitude-frequency correlation in the sunspot cycle

A possible amplitude-frequency correlation (AFC thereafter) in the sunspot cycle, in particular, the importance of the amplitude in determining the length of the related cycle has already been noted in thirties by Waldmeier [177] and recently discussed by Hathaway et al. [178]. In this section we demonstrate that the amplitude-frequency correlation found in the sunspot cycle is probably a non-random phenomenon and propose its explanation by an underlying nonlinear dynamical system.

The series of yearly sunspot numbers from the period 1700 – 2006 (Fig. 15, top panel) has been filtered by a simple moving average (MA) band-pass filter: First, the MA's from a 13-sample window have been subtracted from the data in order to remove slow processes and trends, and then a 3-sample MA smoothing has been used in order to remove high-frequency components and noise. Then the discrete version of the Hilbert transform (36), using the window length of 25 samples, has been applied in order to obtain the instantaneous amplitude $A(t)$ and the instantaneous phase $\phi(t)$. For obtaining a more robust estimation of the instantaneous frequency $\omega(t)$ than the one yielded by a simple differencing the phase $\phi(t)$, the robust linear regression [179] in a 7-sample moving window has been used. Finally, the series of $A(t)$ and $\omega(t)$ have been smoothed using a 13-sample MA window. The resulting series of the instantaneous amplitude and frequency of the yearly sunspot numbers, plotted in Fig. 20a, yield the crosscorrelation equal to 0.505. Does this value mean that the amplitude and frequency of the sunspot cycle are correlated as a consequence of an underlying dynamics, or could this correlation occur by chance? Searching for an answer, we test the statistical significance of this correlation using the surrogate data approach.

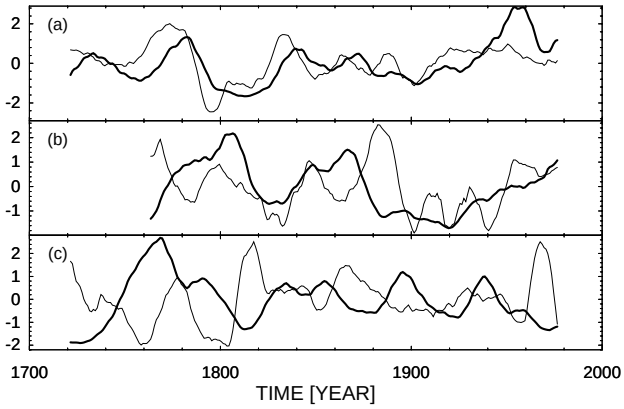

FIG. 20: The instantaneous amplitude (solid line) and frequency (thin line) of (a) the yearly sunspot numbers series, (b) of a realization of the related IAFT surrogate data, and (c) of a realization of the Barnes model.

In the first step we use the FT and IAFT surrogates, introduced in Sec. 4 4.2, where this kind of the surrogate data played the role of a linear stochastic process with the same spectrum and histogram as the studied data. Testing nonlinearity in general, it is stressed that the IAFT surrogates replicate the linear "properties" (more exactly, temporal dependence), while do not contain any nonlinear dependence structure. Here, testing the significance of the AFC, we consider the IAFT surrogates as a data with cycles oscillating with the same frequencies as the sunspot cycles, however, not possessing any systematic amplitude-frequency correlation. Since for generating the IAFT surrogates we use the Fast Fourier Transform (FFT) [179] which requires the number of samples equal to a power of two, we perform two tests, using the "first" and the "last" 256 samples, i.e., the subseries of the whole 307-sample series obtained by cutting away 51 samples at the end, or at the beginning, respectively, from the whole yearly sunspot numbers record. Thus, in each test, the surrogate data replicate the sample spectrum of the related 256-sample subseries.

The FT and IAFT surrogates are generated from the raw (unfiltered) 256-sample segments of the sunspot data. Also, the 256-sample subseries are used for estimating the amplitude-frequency correlation related to the

particular subseries, applying the procedures described above. Then, each realization of the (IA)FT surrogates, generated with respect to the raw data, undergoes the same processing as the raw data, i.e., the MA bandpass filtering, the Hilbert transform and the robust linear regression for the $\omega(t)$ estimation, and the final $A(t)$ and $\omega(t)$ smoothing are performed before computing the AFC for each surrogate realization. Then the *absolute* values of the AFC's for 150,000 surrogate realizations are evaluated in order to assess the significance of the related AFC value found in the sunspot data. The first 256-sample subseries of the sunspot yearly numbers yields the AFC equal to 0.605, while the mean value of the *absolute* AFC for the IAFT surrogate set is 0.26 with the standard deviation (SD) equal to 0.17.

As we already discussed, usually in surrogate tests the significance is derived from the difference between the data value and the surrogate mean, divided by the surrogate SD, provided normal distribution of such defined statistic. Having generated the large amount of the surrogate replications, here we directly estimate the significance p of the test, i.e., the probability that the assessed correlation occurred by chance, considering the chosen null hypothesis (realized by the surrogate model), by simply counting the occurrences in the surrogate set of absolute AFC values greater or equal to the assessed raw data value, i.e., 0.605 in this case. The number obtained is 3637, which is equal to 2.43%. Statistically speaking, the test result is significant on $p < 0.03$, or, in other words, the probability that the amplitude-frequency correlation found in studied segment of the sunspot data occurred by chance (as a random event) is smaller than 3%, given the null hypothesis AFC=0, realized by the used surrogate data.

Processing the 'last' 256-sample segment of the yearly sunspot numbers, the obtained AFC is equal to 0.532, while the values from the IAFT surrogates are the same as above, however, the p-value in this case is 6.58%. Still, we can conclude that the test result is significant on $p < 0.07$. An example of the IAFT surrogate realization is plotted in Fig. 15, bottom panel, its instantaneous amplitude and frequency in Fig. 20b.

The results from the tests using the simple FT surrogates (i.e., without the histogram adjustment) are practically the same as those from the above IAFT surrogates. Testing the monthly sunspot numbers [180] the segments of (the 'first' and the 'last') 2048 samples were used. The same data processing has been applied as described above in the case of the yearly data with the windows lengths equivalent in real time, i.e., multiplied by 12 in number of samples. The obtained results are perfectly equivalent to those yielded by the yearly data, i.e., $p < 0.03$ and $p < 0.07$ for the 'first' and the 'last' 2048-sample segments, respectively.

In the second step of testing the significance of the sunspot cycle AFC, we use realizations of the Barnes model as the surrogate data. As noted above, the Barnes model mimics some important statistical properties of the sunspot data. The distribution of instantaneous frequencies, obtained from realizations of the Barnes model, is consistent with the distribution of the instantaneous frequencies of the real sunspot data [181]. When the Barnes model is considered as the null hypothesis, it is hard to reject it by standard nonlinearity tests, however, realizations of the Barnes model do not possess any systematic amplitude-frequency correlation. A realization of the Barnes model is plotted in Fig. 15, middle panel, its instantaneous amplitude and frequency in Fig. 20c.

In the test, 150,000 realizations of 307-sample series were generated by the Barnes model and processed by the same way as the sunspot series. The mean absolute AFC is equal to 0.21, SD=0.15, comparison with the AFC obtained for the whole 307-sample yearly sunspot series (AFC=0.505) yields the p-value equal to 4.36%. Thus, considering the Barnes model as the null hypothesis, the probability that the sunspot series AFC=0.505 occurred by chance is $p < 0.05$.

Using two different types of stochastic models (iteratively rescaled isospectral surrogates and the Barnes model) which replicate some properties of the sunspot cycle, we have obtained a statistical support for the hypothesis that the amplitude-frequency correlation observed in the sunspot cycle did not occur by chance (as a random event) but is probably a property of an underlying dynamical mechanism. Paluš and Novotná [182] pointed out that well-known systems, possessing this property, are nonlinear oscillators, in which a significant AFC can be observed also in cases of external, even random, driving force. There is, however, no direct connection between the significant AFC and a possible model in a form of a nonlinear oscillator (with possibly random driving) underlying the dynamics of the sunspot cycle. The evidence presented by Paluš and Novotná [182] can be understood as the first step in bridging the gap between reliable statistical analyses of the experimental sunspot data (dominated by linear stochastic methods) and physical models such as nonlinear dynamo models [165] (compared with data only on a qualitative level). Interpreting the results of statistical tests, one should be aware of the precise formulation (and realization) of the null hypothesis and their possible negations. Timmer [183] also pointed out this problem and proposed, in this case, an alternative hypothesis in the form of realizations of the second order linear stochastic (AR2) process with time-variable frequency. Such a process possesses a complex, analytically given relation between its variance and theoretical frequency, which however, is not necessarily reflected in a systematic AFC. Paluš and Novotná [184] demonstrated that such a hypothesis can easily be rejected, so that the nonlinear oscillator is still a more plausible explanation of the observed AFC. A simple nonlinear oscillator model can naturally be considered as a projection from a spatio-temporal field described by nonlinear partial differential equations, i.e., our result does not contradict to magnetohydrodynamic dynamo mod-

els. And indeed, the evidence for the significant AFC in the sunspot data, for the first time presented in [182], inspired Mininni et al. [185] to consider a relaxation oscillator model obtained as a spatial truncation of the magnetohydrodynamic equations. They demonstrated that the AFC in such model data is consistent with the AFC in the sunspot data. Therefore, they require that theoretical models for the solar cycle should be expected to display the significant AFC [185].

7. STATISTICAL TESTING IN THE PROCESS OF SCIENTIFIC DISCOVERY

Let us summarize what we have learned from the above examples. Non-specific tests for nonlinearity can serve well if they bring the negative result – i.e., the null hypothesis is not rejected. It does not mean that the underlying process is indeed linear, however, the quantity used for the discriminating statistic is not able to distinguish the tested data from the null hypothesis of (nonlinearly transformed) linear stochastic process. Then any nonlinearity-specific hypothesis, such as the chaotic attractor, is not supported by the data, in the combination with the used analysis tools. On the other hand, the rejection of the null hypothesis open many possibilities to search for an underlying mechanism. As we have stated above (Sec. 4 4.4), the null hypothesis of a (nonlinearly transformed) linear stochastic process can be rejected by various properties of the data, not necessarily by the functional nonlinearity (11). The positive result of a non-specific test for nonlinearity gives an important information that the standard linear methods for time series analysis might be inadequate, however, is not very informative considering the inference of an actual mechanisms underlying the data. We call the non-specific nonlinearity test the above test with the mutual information and redundancy, since these functionals of probability distributions are general measures of dependence and predictability. In some cases they can serve in specific tests, however, in the context discussed above they did not provide very specific information regarding the mechanism underlying the sunspot cycle. The same holds for 'chaotic' measures such as the correlation dimension or the Lyapunov exponents. These measures are very specific if they were applied to data indeed generated by dynamics on strange attractors, however, they were demonstrated to give finite values also for linear stochastic processes. In such cases the correlation dimension does not reflect the dimensionality of an underlying attractor, and the Lyapunov exponents do not reflect the exponential divergence of trajectories, but these measures reflect some non-specific property of the analysed data. The chaotic measures are usually influenced by the strength and the rate of decay of the autocorrelation function [171]; and also by the amount of noise in the data, sampling rate and other properties specific rather to the measurement than to the underlying dynamical processes.

Therefore, it is more informative to quantify a really existing feature of the data which can point to some known nonlinear property. As an example we choose the amplitude-frequency correlation, a typical feature of nonlinear oscillators. We have demonstrated that such a feature is possessed by the sunspot cycle. As we noted above, some relationship between the amplitude and the frequency or the period of the solar cycles had been observed a long time ago [177, 178], however, Paluš and Novotná [182] showed for the first time that the amplitude-frequency correlation observed in the sunspot cycle did not occur by chance, but it is a property inherent to the dynamics of the sunspot cycle. This finding further inspired the way of choosing among appropriate physical models for the solar cycle [185]. This is the right way how the statistical testing can be helpful in uncovering physical mechanisms underlying complex phenomena: One cannot expect a definitive answer about the nature of the studied process from a single statistical test. A sequence of tests in which both the null hypothesis and the discriminating statistic are developed and targeted to infer specific knowledge about the studied phenomenon can lead to an improvement of our understanding of underlying physical mechanisms and, eventually, in a further step, to a proposal of a valid model with a predictive power. Uncovering of the physical laws underlying complex phenomena is a gradual process in which hypotheses are proposed, appropriately tested and either rejected, or accepted and then potentially refined. An acceptance of an alternative hypothesis is not a definitive decision – a new data, or a new evidence obtained from the known data can falsify a hypothesis and new alternatives should be considered.

8. SOME REMARKS ON NONLINEARITY IN THE HUMAN EEG

Since the 1980s there has been a sustained interest in describing neural processes and brain-signals, especially the electroencephalogram (EEG), within the context of nonlinear dynamics and the theory of deterministic chaos (see, e.g., [48, 186–188] for comprehensive reviews). The electroencephalogram is a record of the oscillations of brain electric potentials recorded from electrodes attached to the human scalp, revealing synaptic action that is moderately to strongly correlated with brain states. If the nature of EEG signals was actually low-dimensional, the results of dimensional analyses could be of immense importance for theoretical neuroscience and for neurological and psychiatric clinical practice. However, confidence of the results such as finite dimensions and positive Lyapunov exponents obtained from EEG data have been considerably challenged by using the surrogate data tests. For instance, in 1989 Rapp et al. [47] have analysed an extensive collection of EEG recordings of healthy subjects at rest and during a mental activity. In all cases

they have found finite correlation dimensions which even discriminated between the resting condition and one type of a mental activity ($D_2 = 4.0 \pm 1.4$ and $D_2 = 4.7 \pm 1.0$, respectively), the difference in a paired t-test was found significant with $p = 0.015$. In 1996 the same data were re-examined by Theiler and Rapp [189] using an improved correlation dimension estimator and the surrogate data tests. The re-examination suggested that the previous indication of low-dimensional structures had been an artifact of autocorrelations in oversampled signals. Signatures of nonlinearity have been found only in a few signals, the vast majority of the analysed EEG records have been found more appropriately modelled by filtered noise (an isospectral linear stochastic process). Moreover, the discriminating power of the correlation dimension have been found worse than the discriminating ability of measures based on the spectral analysis.

Other authors [190, 191] in independent studies have detected a nonlinear component in EEG recordings of normal healthy volunteers, however, signatures of low-dimensional chaos were not found. Paluš [191] analysed two-hour vigilance and sleep EEG recordings from five healthy volunteers, using the surrogate data testing with the redundancy – linear redundancy as the discriminating statistics. With appropriate FT surrogate data well preserving the linear properties (differences in the linear redundancy under 1.5SD), highly significant differences (10–30SD) have been found using the redundancy $R^{(n)}(\tau)$. Several types of nonlinear behaviour, found in the EEG recordings, are described in Ref. [191], here we would like to point to an 'asymmetry' in the plot of $R(y(t); y(t+\tau))$ computed from the EEG data in comparison with the 'symmetric' result obtained from the FT surrogate data (see Fig. 21). It is interesting that a simple simulated signal can reproduce such a behaviour. Let us consider episodes of two-frequency oscillations of the form

$$y(t) = A \sin(\omega_1 t) \sin(\omega_2 t), \tag{38}$$

where $\omega_1 = 2\pi/T_1$ represents fast dominant oscillations (e.g., $T_1 = 0.1$ sec.) and $\omega_2 = 2\pi/T_2$ represents a slower "envelope" (e.g., $T_2 = 0.5$ – 1.0 sec.), randomly distributed in a background of Gaussian noise. If the frequencies $\omega_{1,2}$ randomly fluctuate, the data are consistent with the linear stochastic surrogates. The nonlinear behaviour of the EEG data, depicted in Fig. 21, can be reproduced if the faster frequency ω_1 smoothly changes. It seems that an analysis of the instantaneous frequency of the cyclic EEG activity (e.g., α-spindles) can bring interesting results, similarly as in the above discussed cases of the sunspot numbers or the variable annual cycle in the atmospheric temperature.

We used the term 'asymmetry' in order to characterize the above behaviour of the redundancy $R(y(t); y(t+\tau))$ of the EEG and the simulated data. In the context of time series the notion of temporal symmetry–asymmetry is very interesting. The redundancy R^2 (mutual information I) is not a measure suitable for detecting the

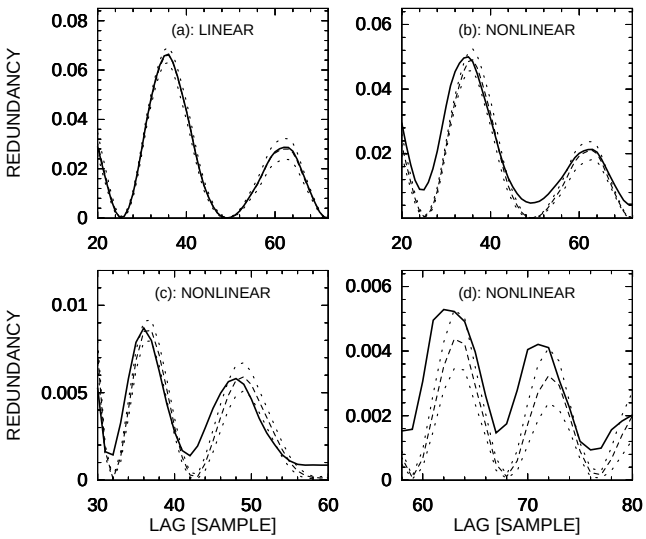

FIG. 21: (a) Linear redundancy $L^2(\tau)$ and (b) redundancy $R^2(\tau)$ for the sleep EEG of Subj. 5 from position P_z (full lines) and its FT surrogates (mean value – dashed lines, mean $\pm$ SD – dotted lines). (c, d) The same as (b), but (c) for the vigilance EEG of Subj. 1, position O_1 and (d) for the simulated data with the smoothly changing fast frequency.

temporal asymmetry, due to its symmetry properties ($I(\xi;\eta) = I(\eta;\xi)$). Considering $\xi = y(t)$ and $\eta = y(t+\tau)$, the temporal asymmetry of a series $\{y(t)\}$ can be detected by the significant difference between the probability distributions $p(\xi,\eta)$ and $p(\eta,\xi)$. The latter can be quantified by the Kullback-Leibler information [137]

$$K(\xi,\eta) = \sum_{\xi}\sum_{\eta}[p(\xi,\eta) - p(\eta,\xi)] \log[p(\xi,\eta)/p(\eta,\xi)].$$

Application of this approach in Ref. [191] led to positive detection of the temporal asymmetry in the EEG. Chialvo & Millonas [192] have pointed out the importance of temporally asymmetric fluctuations for biological energy transduction. Possible existence and a role of temporally asymmetric processes in the brain dynamics could be an interesting subject of research.

Until now we have discussed possible nonlinearity in the dynamics of single EEG signals. The EEG signals ('channels'), however, are usually recorded from a number of electrodes placed in predefined positions on the human scalp. Thus spatio-temporal dynamics of the EEG should be considered [193]. Detection and characterization of interactions between or among the EEG signals from different parts of the scalp is an important and challenging task. Paluš et al. [202, 203] studied interactions in multichannel electroencephalograms of patients suffering from epilepsy. Causal relations between EEG signals measured in different parts of the brain were identified using the approach which will be discussed in the following Sec. 9. In transients from normal brain activity to epileptic seizures, asymmetries in the interactions emerge or are amplified. Nonlinear measures of directional in-

teractions or causality can potentially help to localize an epileptic focus (the part of the brain where seizures start) or even to foresee epileptic seizures [202, 203]. From a rather abstract question of nonlinearity in the dynamics underlying studied data, we will develop the surrogate data testing methodology with information-theoretic discriminating statistics as a tool for detection of directional interactions between signals in the following Sec. 9. The level of interactions, or the quality of information transfer between different parts of the human brain is critical for healthy mental activity. Preliminary results show that the nonlinear measures of interactions can help in early diagnostics of neurodegenerative diseases such as the Alzheimer disease.

9. INFERENCE OF DIRECTIONAL INTERACTIONS OR CAUSALITY IN COMPLEX SYSTEMS

Emergent phenomena in complex systems consisting of many components usually cannot be explained by a simple linear sum or a combination of the properties of the system components. The complex behaviour is usually the result of mutual interactions of the systems components. Knowledge about the way of interactions among the system components can be the key to understanding of the complex phenomena. Among various types of interactions and dependence, synchronization [194] plays a specific role in the cooperative behavior of coupled complex systems or their components. Synchronization and related phenomena have been observed not only in physical, but also in many biological systems. Examples include cardio-respiratory interactions [195–198] and synchronization of neural signals on various levels of organization of brain tissues [199–203]. In such systems it is not only important to detect synchronized states, but also to identify drive-response relationships between studied systems. This task is a special case of the general problem of determining causality or causal relations between systems, processes or phenomena. Wiener [204] proposed mathematical formulation of causality in measurable terms of predictability. Granger [205] introduced a specific notion of causality into time series analysis by evaluation of predictability in bivariate autoregressive models. This linear framework for measuring and testing causality has been widely applied in economy and finance (see Geweke [206] for survey of the literature), but also in diverse fields of natural sciences such as climatology (see [207] and references therein) or neurophysiology, where specific problems of multichannel electroencephalogram recordings were solved by generalizing the Granger causality concept to multivariate cases [208, 209]. However, the limitation of the Granger causality concept to linear relations required further generalizations which emerged especially in the intensively developing field of synchronization of complex systems. Considering the task of identification of drive-response relationships, a multitude of asymmetric dependence measures have been proposed [199, 200, 202, 203, 210–215] and applied in diverse scientific fields such as laser physics [216], climatology [217, 218], cardiovascular physiology [215, 217], neurophysiology [199, 200, 202, 203, 219–221], or finance [222]. In spite of these wide-spread applications of various coupling asymmetry measures, the correct inference of coupling asymmetry, i.e., the identification of the driving and driven systems from experimental time series is far from trivial and requires attentions of theoreticians as well as applied scientists.

In the following we will review some basic considerations and results from Paluš and Vejmelka [223]. We will consider two interacting systems, possibly one of them driving the other. Then the coupling asymmetry, or, as it is called, the directionality of coupling, also identifies causality, or causal relations between the studied systems. The problem of distinguishing the true causality from indirect influences in interactions of three or more systems is beyond the scope of the consideration in this paper.

9.1. Asymmetry in coupling: test systems

We will discuss some issues related to inference of causality (in the above defined sense) from experimental data and the way how statistical testing with the surrogate data can be applied in this problem. For these consideration we will use three numerically generated examples of coupled chaotic systems. As the first example, let us consider the unidirectionally coupled Rössler and Lorenz systems, also studied in Refs. [200, 202, 212], given by the equations

$$\begin{aligned} \dot{x}_1 &= -\alpha\{x_2 + x_3\} \\ \dot{x}_2 &= \alpha\{x_1 + 0.2\, x_2\} \\ \dot{x}_3 &= \alpha\{0.2 + x_3(x_1 - 5.7)\} \end{aligned} \tag{39}$$

for the autonomous Rössler system, and

$$\begin{aligned} \dot{y}_1 &= 10(-y_1 + y_2) \\ \dot{y}_2 &= 28\, y_1 - y_2 - y_1\, y_3 + \epsilon\, x_2^{\beta} \\ \dot{y}_3 &= y_1\, y_2 - \frac{8}{3} y_3 \end{aligned} \tag{40}$$

for the driven Lorenz system in which the equation for $\dot{y}_2$ is augmented by a driving term involving x_2. We will analyse the case with $\alpha = 6$ and $\beta = 2$.

As the second example we will use the unidirectionally coupled Hénon maps, also studied in Refs. [199, 202, 212], defined by the equations

$$\begin{aligned} x_1' &= 1.4 - x_1^2 + b_1\, x_2 \\ x_2' &= x_1 \end{aligned} \tag{41}$$

for the driving system $\{X\}$, and

$$\begin{aligned} y_1' &= 1.4 - (\epsilon\, x_1 y_1 + (1-\epsilon)\, y_1^2) + b_2\, y_2 \\ y_2' &= y_1 \end{aligned} \tag{42}$$

for the response system $\{Y\}$. Here we will study the identical systems with $b_1 = b_2 = 0.3$.

The third example, considered here, will be the unidirectionally coupled Rössler systems given by the equations

$$\begin{aligned} \dot{x}_1 &= -\omega_1 x_2 - x_3 \\ \dot{x}_2 &= \omega_1 x_1 + a_1\, x_2 \\ \dot{x}_3 &= b_1 + x_3(x_1 - c_1) \end{aligned} \tag{43}$$

for the autonomous system, and

$$\begin{aligned} \dot{y}_1 &= -\omega_2 y_2 - y_3 + \epsilon(x_1 - y_1) \\ \dot{y}_2 &= \omega_2 y_1 + a_2\, y_2 \\ \dot{y}_3 &= b_2 + y_3(y_1 - c_2) \end{aligned} \tag{44}$$

for the response system. In the following we will use parameters $a_1 = a_2 = 0.15$, $b_1 = b_2 = 0.2$, $c_1 = c_2 = 10.0$, and frequencies $\omega_1 = 1.015$ and $\omega_2 = 0.985$.

The data from continuous nonlinear dynamical systems were generated by numerical integration based on the adaptive Bulirsch-Stoer method [179] using the sampling interval 0.02617 for the systems (39),(40), and 0.1256 for the systems (43),(44). In the latter case this gives 17 – 21 samples per one period.

In all the cases we denote the driving, autonomous system by $\{X\}$, and the driven, response system by $\{Y\}$. For each of the above three examples we define a set of coupling strength parameter ϵ increasing from $\epsilon = 0$ to an ϵ-value before the synchronization threshold. As Paluš et al. [202] explain, the direction of coupling can be inferred from experimental data only when the studied systems are coupled, but not yet synchronized. Considering numerical examples, the synchronization threshold can be determined using the plot of Lyapunov exponents (LE) of the coupled systems as the function of the coupling strength ϵ. With increasing ϵ, the positive Lyapunov exponent of the driven system (also known as the conditional Lyapunov exponent [224]) decreases and becomes negative just with the ϵ-value giving the synchronization threshold. The plots of the Lyapunov exponents for the Rössler-Lorenz systems (39),(40) can be found in Refs. [202, 212], the LE plots for the coupled Hénon systems (41),(42) in Refs. [199, 202, 212], while further study of the coupled Rössler systems (43),(44), including their LE plot, can be found in Sec. 9 9.4.

For each value of ϵ from the predefined range we numerically generate time series $\{x_i\}$ and $\{y_i\}$ as outputs of the systems $\{X\}$ and $\{Y\}$, obtained by recording the components x_1 and y_1, respectively, and analyse them by using three methods defined in the following Sec. 9 9.2.

9.2. Asymmetry measures

Le Van Quyen et al. [200] use the assumption of existence of a smooth map between trajectories of $\{X\}$ and $\{Y\}$. Their asymmetry measure is based on cross-prediction using the well known idea of mutual neighbors. A state space, known or reconstructed using a time-delay embedding [49] $\mathbf{X}_n = [x_n, x_{n-\tau}, x_{n-2\tau}, \dots]$, must be available. However, instead of using k nearest neighbors, a neighborhood size δ is pre-selected. Considering a map from X to Y, a prediction is made for the value of y_{n+1} one step ahead using the following formula

$$\hat{y}_{n+1} = \frac{1}{|V_\delta(\mathbf{X}_n)|} \sum_{j:\mathbf{X}_j \in V_\delta(\mathbf{X}_n)} y_{j+1}. \tag{45}$$

The volume $V_\delta(\mathbf{X}_n) = \{\mathbf{X}_{n'} : |\mathbf{X}_{n'} - \mathbf{X}_n| < \delta\}$ is the δ-neighborhood of $\mathbf{X}_n$ and $|V_\delta(\mathbf{X}_n)|$ denotes the number of points contained in the neighborhood. Using data rescaled to the zero mean and the unit variance, Le Van Quyen et al. [200] define a crosspredictability index by subtracting the root-mean-square prediction error from one

$$P(X \to Y) = 1 - \sqrt{\frac{1}{N}\sum_{n=1}^{N}(\hat{y}_{n+1} - y_{n+1})^2}, \tag{46}$$

which should measure how the system $\{X\}$ influences the evolution of the system $\{Y\}$. The crosspredictability index $P(Y \to X)$ in the opposite direction, characterizing the ability of the system $\{Y\}$ to influence the evolution of the system $\{X\}$ is defined in full analogy.

The second approach, proposed by Arnhold et al. [210] and discussed by Quian Quiroga et al. [212], also explore the assumption of the existence of a smooth map between the trajectories of $\{X\}$ and $\{Y\}$. If such a smooth map exists then closeness of points in the state space X of the system $\{X\}$ implies a closeness of points in the state space Y of the system $\{Y\}$. Therefore mean square distances are used instead of the cross-predictions in order to quantify the closeness of points in both spaces. We use the implementation according to Ref. [212, 226] in which a time-delay embedding [49] is first constructed in order to obtain state space vectors $\mathbf{X}$ and $\mathbf{Y}$ for both time series $\{x_i\}$ and $\{y_i\}$, respectively, then the mean squared distance to k nearest neighbors is defined for each $\mathbf{X}$ as

$$R_n^{(k)}(\mathbf{X}) = \frac{1}{k}\sum_{j=1}^{k} |\mathbf{X}_n - \mathbf{X}_{r_{n,j}}|^2, \tag{47}$$

where $r_{n,j}$ denotes the index of the $j-$th nearest neighbor of $\mathbf{X}_n$. The $\mathbf{Y}$-conditioned squared mean distance is defined by replacing the nearest neighbors of $\mathbf{X}_n$ by the equal time partners of the nearest neighbors of $\mathbf{Y}_n$ as

$$R_n^{(k)}(\mathbf{X}|\mathbf{Y}) = \frac{1}{k}\sum_{j=1}^{k} |\mathbf{X}_n - \mathbf{X}_{s_{n,j}}|^2, \tag{48}$$

where $s_{n,j}$ denotes the index of the $j-$th nearest neighbor of $\mathbf{Y}_n$. Then the asymmetric measure

$$S^{(k)}(\mathbf{X}|\mathbf{Y}) = \frac{1}{N}\sum_{n=1}^{N} \frac{R_n^{(k)}(\mathbf{X})}{R_n^{(k)}(\mathbf{X}|\mathbf{Y})}. \tag{49}$$

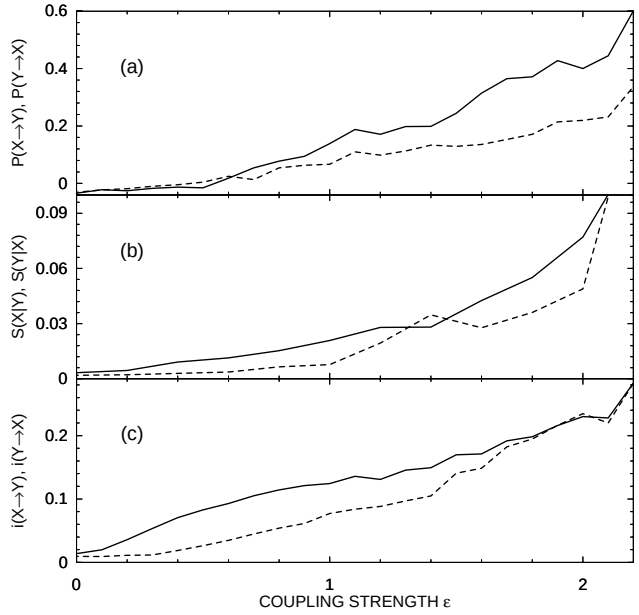

FIG. 22: (a) Cross-predictability $P(X \to Y)$ (solid line) and $P(Y \to X)$ (dashed line), (b) relative average distance of the mutual nearest neighbours $S^{(k)}(\mathbf{Y}|\mathbf{X})$ (solid line) and $S^{(k)}(\mathbf{X}|\mathbf{Y})$ (dashed line), and (c) coarse-grained transinformation rate $i(X \to Y)$ (solid line) and $i(Y \to X)$ (dashed line) for the Rössler system (39) driving the Lorenz system (40), as functions of the coupling strength ϵ.

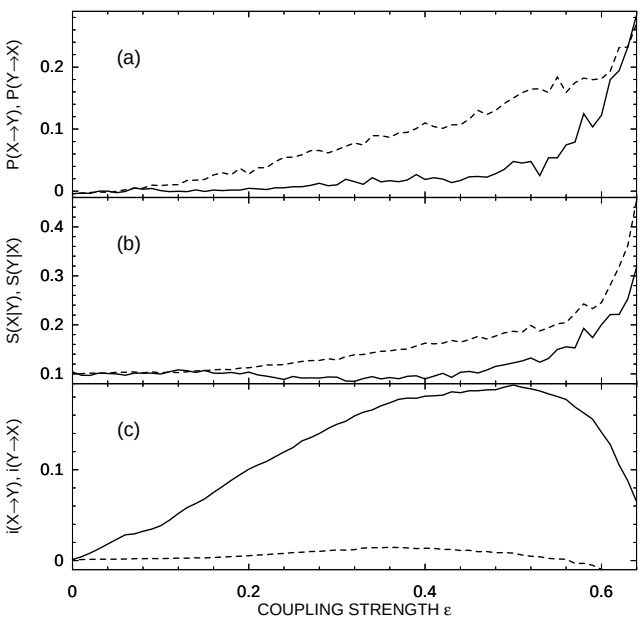

FIG. 23: (a) Cross-predictability $P(X \to Y)$ (solid line) and $P(Y \to X)$ (dashed line), (b) relative average distance of the mutual nearest neighbours $S^{(k)}(\mathbf{Y}|\mathbf{X})$ (solid line) and $S^{(k)}(\mathbf{X}|\mathbf{Y})$ (dashed line), and (c) coarse-grained transinformation rate $i(X \to Y)$ (solid line) and $i(Y \to X)$ (dashed line) for the unidirectionally coupled Hénon system (41),(42), as functions of the coupling strength ϵ.

is supposed to reflect interdependence in the sense that closeness of the points in Y implies closeness of their equal time partners in X and the values of $S^{(k)}(\mathbf{X}|\mathbf{Y})$ approach to one, while, in the case of X independent of Y, $S^{(k)}(\mathbf{X}|\mathbf{Y}) \ll 1$. The quantity $S^{(k)}(\mathbf{Y}|\mathbf{X})$ measuring the influence of $\{X\}$ on $\{Y\}$ is defined in full analogy.

The third measure, used in the following three examples, the coarse-grained transinformation rate $i(X \to Y)$ is the average rate of the net amount of information "transferred" from the process $\{X\}$ to the process $\{Y\}$, or, in other words, the average rate of the net information flow by which the process $\{X\}$ influences the process $\{Y\}$. The coarse-grained transinformation rate (CTIR), introduced by Paluš et al. [202], is based on the conditional mutual information defined in Sec. 4 4.3 and further developed in Sec. 9 9.4. Here we only indicate that the CTIR $i(X \to Y)$ is given by the conditional mutual information $I(x(t); y(t+\tau)|y(t))$ averaged over a range of time lags τ; and analogously, $i(Y \to X)$ is given by τ-averaged $I(y(t); x(t+\tau)|x(t))$. The symmetric term $I(x(t); y(t+\tau))$, averaged over a range of negative and positive τ's, is subtracted in the both cases [202].

9.3. Asymmetric measures and causality

The three asymmetric measures, introduced above, as functions of the coupling strength ϵ, for the Rössler system (39) driving the Lorenz system (40) are plotted in Fig. 22. Except for the weakest coupling ($\epsilon \leq 0.6$), the cross-predictability of the system $\{Y\}$ by the system $\{X\}$ (the solid line in Fig. 22a) is greater than the cross-predictability of the system $\{X\}$ by the system $\{Y\}$ (the dashed line in Fig. 22a). Our result in Fig. 22a agrees with that of Le Van Quyen et al. [200] who interpret the relation $P(X \to Y) > P(Y \to X)$ by the fact, that the autonomous Rössler system $\{X\}$ drives the response Lorenz system $\{Y\}$ and therefore the prediction of $\{Y\}$ from $\{X\}$ is better than the prediction in the opposite direction.

In a similar way, with a few exemptions, the relative average distance of the mutual nearest neighbours $S^{(k)}(\mathbf{Y}|\mathbf{X}) > S^{(k)}(\mathbf{X}|\mathbf{Y})$ (Fig. 22b) agrees with the results in [212], suggesting that the state of the driven system $\{Y\}$ depends more on the state of the driver system $\{X\}$ than vice versa, as also claimed by Arnhold et al. [210]. (Note that the conditioning $X|Y$ reflects the influence $Y \to X$ and vice versa.) The same conclusion about $\{X\}$ driving $\{Y\}$ can be drawn from the CTIR $i(X \to Y) > i(Y \to X)$ (Fig. 22c). The latter inequality holds for all positive values of ϵ but the ϵ-values approaching the synchronization threshold which emerges for ϵ slightly exceeding 2 [202, 212].

The same analyses as presented in Fig. 22, but for the unidirectionally coupled Hénon system (41),(42), are shown in Fig. 23. One can immediately see that in this case $P(X \to Y) < P(Y \to X)$ (Fig. 23a). This result agrees with that of Ref. [199]. Shiff et al. [199] offer an

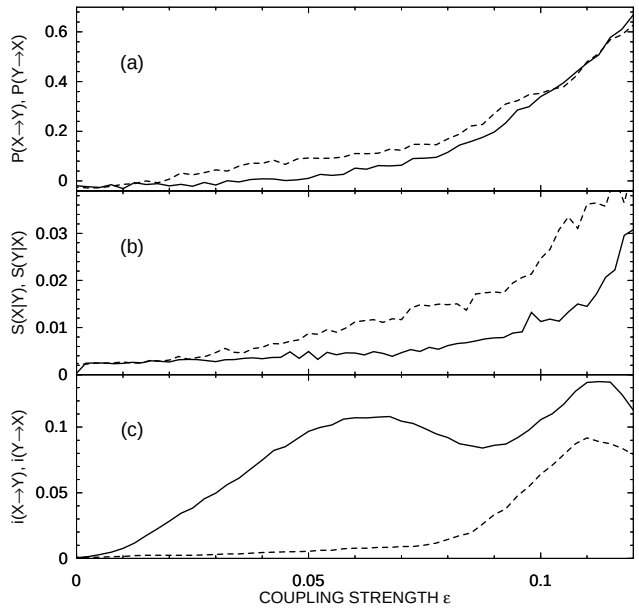

FIG. 24: (a) Cross-predictability $P(X \to Y)$ (solid line) and $P(Y \to X)$ (dashed line), (b) relative average distance of the mutual nearest neighbours $S^{(k)}(\mathbf{Y}|\mathbf{X})$ (solid line) and $S^{(k)}(\mathbf{X}|\mathbf{Y})$ (dashed line), and (c) coarse-grained transinformation rate $i(X \to Y)$ (solid line) and $i(Y \to X)$ (dashed line) for the unidirectionally coupled Rössler systems (43),(44), as functions of the coupling strength ϵ.

interpretation based on the Takens embedding theorem [49]: From the time series $\{x_i\}$ only the system $\{X\}$ can be reconstructed, while from the time series $\{y_i\}$ the whole system consisting of the coupled systems $\{X\}$ and $\{Y\}$ can be reconstructed and therefore one can predict the driving system from the response system and not vice versa [199]. Also, the relation of the second interdependence measure reverses: In this case the inequality $S^{(k)}(\mathbf{Y}|\mathbf{X}) < S^{(k)}(\mathbf{X}|\mathbf{Y})$ holds. (Fig. 23b). Again, our result agrees with that of Ref. [212]. Quian Quiroga et al. [212] explain that the higher-dimensional system (obtained by the reconstruction from the time series $\{y_i\}$ which bears information about both the coupled systems) is 'more active' than the lower-dimensional (autonomous, driving) system. Only the CTIR gives the same relation as in the previous case: $i(X \to Y) > i(Y \to X)$ (Fig. 23c) suggesting the fact that $\{Y\}$ is influenced by $\{X\}$, while $\{X\}$ evolves autonomously.

The analysis of the unidirectionally coupled Rössler systems (43),(44) is presented in Fig. 24. We can see that the results are in a qualitative agreement with those of the coupled Hénon systems (Fig. 23), although these systems are more similar to the first example of the coupled Rössler-Lorenz systems. We can se that neither the cross-predictability, nor the mutual nearest neighbours statistics give consistent results when using three different examples of unidirectionally coupled systems. Only the coarse-grained transinformation rate correctly identifies the direction of the causal influence in the above three examples as well as in many other systems of different origins.

In the above discussed examples of unidirectionally coupled systems we could see that the used measures are generally non-zero in both directions even before the systems become synchronized and comparison of the values of such measures does not always reflect the true causality given by the unidirectional coupling of the studied systems. The intuitively understandable implication *lower prediction error (better predictability)* $\Rightarrow$ *stronger dependence* cannot generally be applied for nonlinear systems. When the coupling of systems is weaker than that necessary for the emergence of synchronization, any smooth deterministic function between the states of the systems cannot exist yet. However, there is already some statistical relation valid on the coarse-grained description level. Although the deterministic quantities are based on the existence of a smooth functional relation, when estimated with finite precision, they usually give nonzero values influenced not only by the existing statistical dependence, but also by properties of the systems other then the coupling. It is therefore necessary to use quantities suitable for measuring statistical dependence, such as information-theoretic measures, which have solid mathematical background and their properties have thoroughly been studied since their introduction in 1948 [135].

9.4. Inference of causality with the conditional mutual information

Let $\{x(t)\}$ and $\{y(t)\}$ be time series considered as realizations of stationary and ergodic stochastic processes $\{X(t)\}$ and $\{Y(t)\}$, respectively, $t = 1, 2, 3, \ldots$. In the following we will mark $x(t)$ as x and $x(t+\tau)$ as x_τ, and the same notation holds for the series $\{y(t)\}$.

The mutual information $I(y; x_\tau)$ measures the average amount of information contained in the process $\{Y\}$ about the process $\{X\}$ in its future τ time units ahead (τ-future thereafter). This measure, however, could also contain an information about the τ-future of the process $\{X\}$ contained in this process itself, if the processes $\{X\}$ and $\{Y\}$ are not independent, i.e., if $I(x; y) > 0$. In order to obtain the "net" information about the τ-future of the process $\{X\}$ contained in the process $\{Y\}$ we need the conditional mutual information $I(y; x_\tau|x)$. The latter was used by Paluš et al. [202] to define the coarse-grained transinformation rate, used in the above three examples. We used the standard statistical language in which we considered the time series $\{x(t)\}$ and $\{y(t)\}$ as realizations of stochastic processes $\{X(t)\}$ and $\{Y(t)\}$, respectively. If the processes $\{X(t)\}$ and $\{Y(t)\}$ are substituted by dynamical systems evolving in measurable spaces of dimensions m and n, respectively, the variables x and y in $I(y; x_\tau|x)$ and $I(x; y_\tau|y)$ should be considered as $n-$ and $m-$dimensional vectors. In experimental practice, however, usually only one observable is recorded for each system. Therefore, instead of the original components

of the vectors $\vec{X}(t)$ and $\vec{Y}(t)$, the time delay embedding vectors according to Takens [49] are used. Then, back in time-series representation, we have

$$I\big(\vec{Y}(t);\vec{X}(t+\tau)|\vec{X}(t)\big) = I\Big(\big(y(t),y(t-\rho),\ldots,y(t-(m-1)\rho)\big);x(t+\tau)|\big(x(t),x(t-\eta),\ldots,x(t-(n-1)\eta)\big)\Big), \quad (50)$$

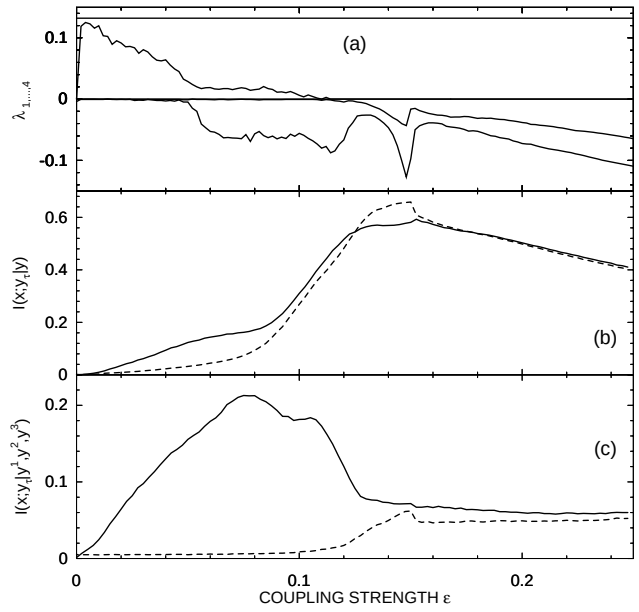

FIG. 25: (a) The two largest Lyapunov exponents of the drive $\{X\}$ (constant lines) and of the response $\{Y\}$ (decreasing lines), (b) averaged conditional mutual information $I(x;y_\tau|y)$ (solid line) and $I(y;x_\tau|x)$ (dashed line), and averaged CMI $I(x;y_\tau|\vec{Y})$ (solid line) and $I(y;x_\tau|\vec{X})$ (dashed line), where the vectors $\vec{X}$ and $\vec{Y}$ are the original components of the integrated systems, for the unidirectionally coupled Rössler systems (43),(44), as functions of the coupling strength ϵ.

where η and ρ are time lags used for the embedding of trajectories $\vec{X}(t)$ and $\vec{Y}(t)$, respectively. Only information about one component $x(t+\tau)$ in the τ-future of the system $\{X\}$ is used for simplicity. The opposite CMI $I\big(\vec{X}(t);\vec{Y}(t+\tau)|\vec{Y}(t)\big)$ is defined in full analogy. Exactly the same formulation can be used for Markov processes of finite orders m and n. Based on the idea of finite-order Markov processes, Schreiber [211] has proposed a "transfer entropy" which is in fact an equivalent expression for the conditional mutual information (50) – see Refs. [137, 223].

Let us consider again the unidirectionally coupled Rössler systems (43),(44). The dependence of their Lyapunov exponents (all but the two which are negative in the uncoupled case) on the coupling strength ϵ is plotted in Fig. 25a. The change of the positive LE of the response system $\{Y\}$ to negative values with the coupling strength slightly under $\epsilon = 0.12$ gives the synchronization threshold for these systems. If we evaluate the simple CMI $I(y;x_\tau|x)$ and $I(x;y_\tau|y)$ with one-dimensional condition x or y, the CMI's in both direction are positive and increasing with the increasing coupling strength (Fig. 25b). Before the synchronization threshold, the inequality $I(x;y_\tau|y) > I(y;x_\tau|x)$ indicates the correct direction of coupling, however, as we will see in the next Section 9 9.5, for reliable inference in general, it is desirable to obtain a zero value in the uncoupled direction $Y \to X$. This can be attained by a proper conditioning – the conditioning variable should contain full information about future values of the system or process generating this variable in the uncoupled case. So it should be three-dimensional vectors $\vec{X}$ or $\vec{Y}$ for the studied Rössler systems. On the other hand, it is sufficient to have just one component of each vector variable for establishing the presence of coupling, i.e., the appropriate measures for inference of coupling directions are the CMI's $I(x;y_\tau|\vec{Y})$ and $I(y;x_\tau|\vec{X})$. Evaluation of the latter quantities brings a five-dimensional estimation problem which might be hard to solve with limited amount of available data.

The CMI's $I(x;y_\tau|\vec{Y})$ and $I(y;x_\tau|\vec{X})$ obtained using the original components $x_1(t)$, $x_2(t)$, $x_3(t)$ and $y_1(t)$, $y_2(t)$, $y_3(t)$ for the conditioning vectors $\vec{X}$ and $\vec{Y}$, respectively, are displayed in Fig. 25c. We can see that $I(y;x_\tau|\vec{X})$ in the uncoupled direction stays at the zero value up to ϵ close to the synchronization threshold, while $I(x;y_\tau|\vec{Y})$ is distinctly positive (Fig. 25c).

The CMI's $I(x;y_\tau|\vec{Y})$ and $I(y;x_\tau|\vec{X})$ with the conditioning vectors $\vec{X}$ and $\vec{Y}$ obtained as the time-delay embedding [49] from the components $x_1(t)$ and $y_1(t)$, respectively, are displayed in Fig. 26a. We can see entirely equivalent results in Fig. 25c and in Fig. 26a.

Many interesting processes in physics, biology, or technology can be modelled by weakly coupled oscillators and their interactions can be inferred by analyzing the dynamics of their instantaneous phases [194, 213, 214]. The instantaneous phase $\phi(t)$ from a time series can be obtained according to Eq. (37) in Sec. 6 6.1.

Paluš & Stefanovska [215] have shown that the conditional mutual information can be applied also in inference of coupling of systems using their instantaneous phases, confined in interval $[0, 2\pi)$ or $[-\pi, \pi)$ (so-called wrapped phases). Thus we can come back to the time series $\{x(t)\}$ and $\{y(t)\}$ generated by the unidirectionally coupled Rössler systems (43),(44) and compute their

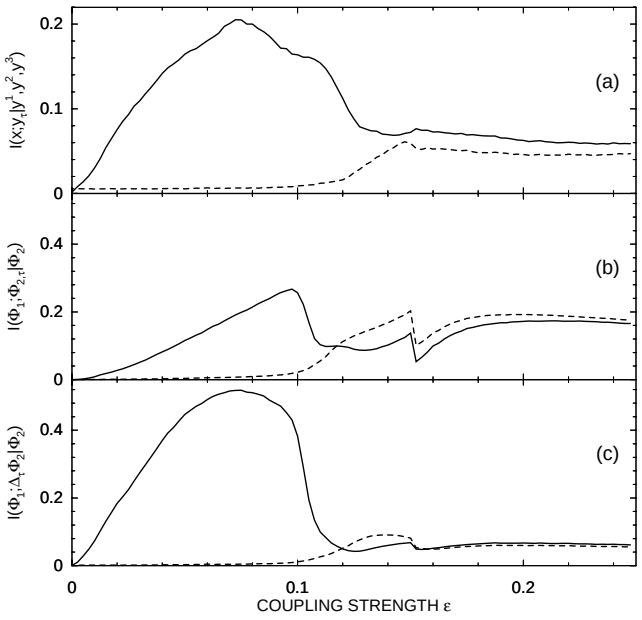

FIG. 26: (a) Averaged conditional mutual information $I(x; y_\tau|\vec{Y})$ (solid line) and $I(y; x_\tau|\vec{X})$ (dashed line), using the time-delay embedding vectors $\vec{X} = [x_1(t), x_1(t-\eta), x_1(t-2\eta)]$ and analogously for $\vec{Y}$, (b) averaged CMI $I(\phi_1(t); \phi_2(t+\tau)|\phi_2(t))$ (solid line) and $I(\phi_2(t); \phi_1(t+\tau)|\phi_1(t))$ (dashed line), and (c) averaged CMI $I(\phi_1; \Delta_\tau\phi_2|\phi_2)$ (solid line) and $I(\phi_2; \Delta_\tau\phi_1|\phi_1)$ (dashed line), for the unidirectionally coupled Rössler systems (43),(44), as functions of the coupling strength ϵ.

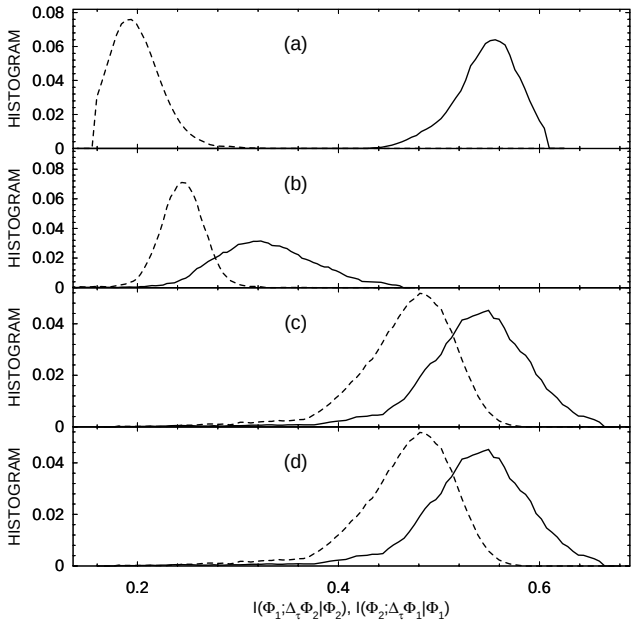

FIG. 27: Histograms of estimates of $I(\phi_1; \Delta_\tau\phi_2|\phi_2)$ (solid lines) and $I(\phi_2; \Delta_\tau\phi_1|\phi_1)$ (dashed lines) from 1000 realizations of (a) the unidirectionally coupled Rössler systems (43),(44), the coupling strength $\epsilon = 0.05$ and the number of samples $N = 1024$, (b) FFT surrogate data for the data used in (a), $N = 1024$; (c) the same as in (a), but the number of samples $N = 512$, and (d) FFT surrogate data for the data used in (c), $N = 512$.

instantaneous phases $\phi_1(t)$ and $\phi_2(t)$, respectively, according to Eqs. (36) and (37). Then we evaluate the conditional mutual information $I(\phi_1(t); \phi_2(t+\tau)|\phi_2(t))$ and $I(\phi_2(t); \phi_1(t+\tau)|\phi_1(t))$ and plot the results in Fig. 26b. We can see that CMI evaluated from the phases distinguishes the driving from the driven system. Moreover, the application of the phase dynamics decreases the dimensionality of the problem – already $I(\phi_2(t); \phi_1(t+\tau)|\phi_1(t))$ with the one-dimensional condition is zero in the uncoupled direction. Even better distinction (Fig. 26c) can be obtained when we study dependence between the phase of one system and the phase increment

$$\Delta_\tau\phi_{1,2}(t) = \phi_{1,2}(t+\tau) - \phi_{1,2}(t), \qquad (51)$$

of the second system instead of the dependence between $\phi_{1,2}(t)$ and $\phi_{2,1}(t+\tau)$. Therefore we evaluate the conditional mutual information $I(\phi_1(t); \Delta_\tau\phi_2(t)|\phi_2(t))$ and $I(\phi_2(t); \Delta_\tau\phi_1(t)|\phi_1(t))$, in a shorter notation $I(\phi_1; \Delta_\tau\phi_2|\phi_2)$ and $I(\phi_2; \Delta_\tau\phi_1|\phi_1)$, respectively.

9.5. Conditional mutual information as a discriminating statistic

Every quantity, descriptive of a state of a system or a process under study, suffers from bias and variance when estimated from noisy, nonstationary experimental data. Using limited, relatively short time series, estimates of complicated quantities such as the conditional mutual information can have non-negligible bias and variance even if evaluated from noise-free, stationary model data. It is necessary to know the behaviour of the used estimator of any measure before it is applied in analysis of real data. In order to study the bias and variance of the CMI estimates, we choose a particular coupling strength ($\epsilon = 0.05$) and evaluate $I(\phi_1; \Delta_\tau\phi_2|\phi_2)$ and $I(\phi_2; \Delta_\tau\phi_1|\phi_1)$ from 1000 realizations of the unidirectionally coupled Rössler systems (43),(44) starting in different initial conditions, for various time series lengths. We evaluate the CMI using a simple box-counting algorithm based on marginal equiquantization, described in Sec. 4 4.3 above.

Histograms obtained from the 1000 CMI estimates, using the time series length $N = 1024$ samples, are plotted in Fig. 27a. We can see the relatively large variance of the estimates and the clear bias of $I(\phi_2; \Delta_\tau\phi_1|\phi_1)$ in the uncoupled direction, however, the distinction between the coupled and the uncoupled directions is still clear. When we use the time series length $N = 512$ samples (Fig. 27c), the variance increases, and the bias in the uncoupled direction rises so that the values of $I(\phi_1; \Delta_\tau\phi_2|\phi_2)$ (solid lines) and $I(\phi_2; \Delta_\tau\phi_1|\phi_1)$ (dashed lines) partially overlap. It is clear that we need some statistical approach to establish critical values of the CMI estimates from which

we could infer that the CMI is nonzero due to a coupling and not due to the estimator bias. In other words, we need to find out what bias and variance we can expect from our data, if there is no coupling. And this is again the task for statistical testing using the surrogate data. Paluš and Vejmelka [223] show that the bias in CMI estimates depends on relative complexity of the dynamics in the sense that the CMI are biased in the direction from the less complex (e.g., periodic) to the more complex (e.g., chaotic) systems. When the systems have a comparable complexity, the bias is influenced by the ratio of dominant frequencies of the studied systems, in the sense that the CMI are biased in the direction from the slower to the faster system. Therefore, in the surrogate data we need to preserve the frequency content of the original data. For this purpose we use the well-known FT surrogate data, introduced in Sec. 4 4.2. Since we use the equiquantal estimator of the conditional mutual information which is independent of the marginal distribution of the data, we do not use any histogram transformation. The surrogate data are constructed from the original time series and then the phases ϕ_1, ϕ_2 are computed from each surrogate realization in the same way as from the original data.

Studying interactions between processes, various null hypotheses can be specified which are reflected in various types of the uni-, bi- or multivariate surrogate data [227]. Our null hypothesis is independence between the phases ϕ_1, ϕ_2. Therefore we construct univariate FT surrogate data for each time series, i.e., the two time series $\{x(t)\}$ and $\{y(t)\}$ are randomized independently. If a specifically nonlinear causality is of interest, then the bivariate surrogate data [228, 229] can be constructed, which preserve cross-correlation functions in addition to the autocorrelation functions.

Histograms of estimates of $I(\phi_1; \Delta_\tau\phi_2|\phi_2)$ (solid lines) and $I(\phi_2; \Delta_\tau\phi_1|\phi_1)$ (dashed lines) from the FFT surrogate data using series lengths $N = 1024$ and $N = 512$ samples are plotted in Fig. 27b and Fig. 27d, respectively. We can see that the average bias of the CMI $I(\phi_2; \Delta_\tau\phi_1|\phi_1)$ in the direction $Y \to X$ in the surrogate data is a bit larger than in the original data (cf. Fig. 27a and Fig. 27b). This fact helps us to avoid false detections of causality (positive information flow) in the uncoupled direction: Even though $I(\phi_2; \Delta_\tau\phi_1|\phi_1)$ from the data gains positive values, these values are not greater than the values from the (uncoupled) surrogates and thus such positive CMI values cannot be considered as the evidence for causality, nor for a directional interaction. In order to translate these considerations into a statistical test, we integrate the histogram of the surrogate CMI values into a cumulative histogram and find the CMI critical values giving 95% of the CMI distribution, counting from the left side. If a CMI value from the tested data is greater than this critical value, the result is significant at the level $\alpha = 0.05$ and the null hypothesis of independence is rejected. Here we employ the Neyman-Pearson approach of hypothesis testing since we do not infer causality from

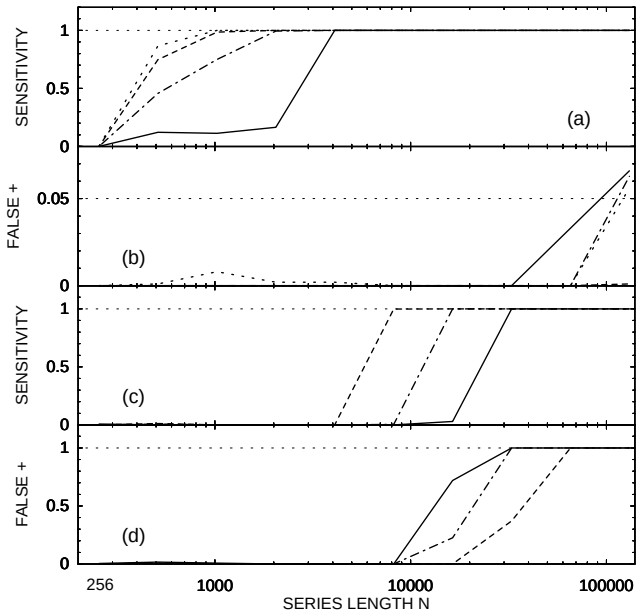

FIG. 28: Sensitivity (a,c) and the rate of false positives (b,d) as functions of time series length N for the tests using the coupled Rössler systems (43),(44) with the frequency ratio 1:1 for different amounts of noise in the data. The portions of noise are in (a,b) 0% (dotted line), 10% (dashed line), 20% (dash-and-dotted line) and 30% (solid line); in (c,d) 50% (dashed line), 70% (dash-and-dotted line) and 100% (solid line).

a single experimental data. Using the cumulative histograms obtained from the 1000 surrogate realizations and having set the nominal value for the significance, $\alpha = 0.05$, leading to a critical value for each test, we use the 1000 realizations of the data from the Rössler systems (43),(44) for the evaluation of the performance of the test. Comparing the values of $I(\phi_2; \Delta_\tau\phi_1|\phi_1)$ in the uncoupled direction with their critical values we obtain the rate of false positive results (type I error, see Sec. 3 3.1), while using $I(\phi_1; \Delta_\tau\phi_2|\phi_2)$ in the coupled, causal direction we count the rate of the correctly positive results, so that we evaluate the sensitivity of the test. If we write the sensitivity as $1 - \beta$, then β is the type II error.

The distribution of $I(\phi_1; \Delta_\tau\phi_2|\phi_2)$ from the surrogate data with $N = 1024$ (Fig. 27b, solid line) allows 100% sensitivity, i.e. values of $I(\phi_1; \Delta_\tau\phi_2|\phi_2)$ from all 1000 realizations of the original Rössler time series were correctly detected as significant, reflecting truly nonzero causal information flow from $\{X\}$ to $\{Y\}$. In the opposite direction (cf. the histograms plotted by the dashed lines in Fig. 27a and Fig. 27b) we have got eight false detections from 1000 realization, i.e. the false detection rate is 0.008, still well under the nominal $\alpha = 0.05$. Using $N = 512$ samples the sensitivity is worse, giving the value 0.866, i.e. 134 realizations from 1000 were not recognized by the test. The false detection rate, however, was 0.001. With an insufficient amount of data the sensitivity of the test could be lowered, however, the surrogate test prevents false detections very well. The dependence

of the sensitivity as well as of the rate of false positives (actual type I error) on the series length is plotted, using the dotted line, in Fig. 28a, and Fig. 28b, respectively. As stated above, 100% sensitivity (actual type II error $\beta = 0$) is reached from the time series length $N = 1024$ samples. With $N = 512$ samples, the actual type II error $\beta = 0.134$, and for shorter data $\beta \to 1$ (Fig. 28a, dotted line), i.e., the test looses its sensitivity. On the other hand, the actual type I error is well under the nominal $\alpha = 0.05$ (Fig. 28b, dotted line), so that the false detections of causality do not occur. In other words, a properly done surrogate data test prevents from the false detection of causality, however, in order to safely detect the true causality we need, at least, about a thousand samples of properly sampled data [223]. This result was obtained using the chaotic oscillators of close frequencies. If the frequencies of the two systems are different, even more data are necessary in order to have sensitive tests for causality [223].

Considering applications to real data, influence of noise on the presented test should also be evaluated. Gaussian noise has been added to the raw data from the coupled Rössler systems (43),(44) with the frequency ratio 1:1 (exactly 1.015:0.985). The amount of added noise is characterized by the noise standard deviation (SD) expressed in the percentage of the SD of the original data, e.g., 10% of noise means SD(noise)=0.1SD(data), or 100% of noise means SD(noise)=SD(data). The noised data were processed and tested in the same way as the noise-free data above. The test sensitivity, i.e., the rate of true positive detections of causality, as well as the rate of false positives, i.e., the rate of formal detections of causality in uncoupled directions, as functions of time series length N are illustrated in Fig. 28. The higher the amount of noise in the data, the more data samples are required in order to obtain 100% sensitivity of the test (Fig. 28a). For moderate amounts of noise, the rate of false positives remains well under or about the nominal test critical value $\alpha = 0.05$ (Fig. 28b). With large amounts of noise, however, the attainment of the 100% sensitivity is followed by an increase of the rate of false positives. With 100% of noise in the data, the rate of false positives goes to 1 (i.e., to 100%) even before the sensitivity rises from 0 to 1, i.e., the detection ability of the test is completely lost. For amounts from 50% of noise there is a bounded range of time series lengths for which the test is reliable, e.g., from 8k to 32k samples for 50%, but only around 16k samples for 70% of noise. The applicability of the test is limited when the data are contaminated by a large amount of noise.

9.6. Testing the direction of the cardiorespiratory interaction

In order to demonstrate how the discussed approach can be applied to real data, we use cardiac and respiratory data from an animal experiment described in

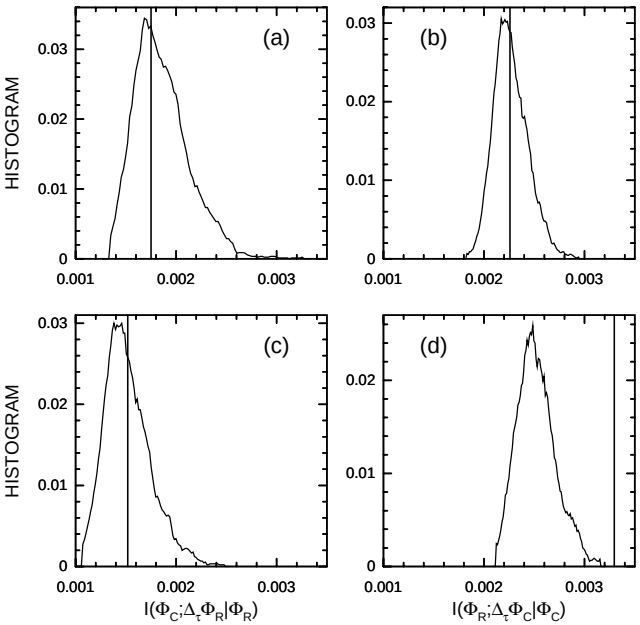

FIG. 29: Tests of causal influences of cardiac oscillations on respiratory oscillations measured by $I(\phi_C; \Delta_\tau \phi_R | \phi_R)$ (a, c) and of the influence of the respiratory rhythm on the cardiac oscillations given by $I(\phi_R; \Delta_\tau \phi_C | \phi_C)$ (b, d). Values from the tested data are marked by the vertical lines, the surrogate ranges are illustrated by the histograms obtained from 2500 surrogate realizations. (a,b) The test of bias using data from two animals; (c,d) a real test for one of the animals.

[198]. Detailed account of the causality analysis in the cardio-respiratory interactions of anaesthetized rats can be found in Ref. [230]. Using the inter-beat and inter-breath intervals we construct so-called marked events phases ϕ_C, ϕ_R for the cardiac and respiratory dynamics, respectively. Let t_k and t_{k+1} be the times of two consecutive events, here peaks in the signal (ECG – electrocardiogram, or the respiratory signal). The instantaneous phases are then linearly interpolated as [198]:

$$\phi(t) = 2\pi \frac{t - t_k}{t_{k+1} - t_k}, \quad t_k \le t < t_{k+1}. \tag{52}$$

Considering, in applications of the marked events phases, that the only available data are the event times $\ldots, t_k, t_{k+1}, \ldots$, or the inter-event intervals $\ldots, t_{k+1} - t_k, \ldots$, a simple way to construct surrogate data is a random permutation of the inter-event intervals before the surrogate marked events phases are computed according to (52).

We estimated $I(\phi_C; \Delta_\tau \phi_R | \phi_R)$ and $I(\phi_R; \Delta_\tau \phi_C | \phi_C)$ from 33-minute recordings which gave, after subsampling to 40 Hz, the series length $N = 80,000$ samples. Thus we can expect a good performance of the tests using the marked events phases from experimental, possibly noisy data. In the first test, presented in Figs. 29a,b, we tried to evaluate the bias of the CMI estimator and the ability of the surrogate data test to prevent possible false detections. For this test we used the cardiac data from one

animal and the respiratory data from another animal, so no true causality cannot exist in this case. The CMI estimates are positive (although small, but this is typical using the marked events phases [223]) in both the directions and the value of $I(\phi_R; \Delta_\tau \phi_C | \phi_C)$ (Fig. 29b) reflecting the influence of the (slower) respiratory rhythm on the (faster) cardiac dynamics is larger than $I(\phi_C; \Delta_\tau \phi_R | \phi_R)$ (Fig. 29a) in the opposite direction. Both values from the tested data, however, lie well inside the surrogate histograms. The latter result means that the CMI values are not significantly larger than zero and no causality, or no information flow exists in the either direction, as we expected using the data from two different animals. The situation is different when analyzing data from a single animal. While there is no significant influence of the cardiac dynamics on the respiration (Fig. 29c), the influence of the respiratory rhythm on the cardiac dynamics is clearly significant, since $I(\phi_R; \Delta_\tau \phi_C | \phi_C)$ from the tested data lies outside the surrogate $I(\phi_R; \Delta_\tau \phi_C | \phi_C)$ distribution (Fig. 29d) and thus even this small value $I(\phi_R; \Delta_\tau \phi_C | \phi_C) = 0.0033$ nat is significantly positive with $p = 0$. The null hypothesis of independence, realized by the randomly permutated heartbeats and respiratory cycles, was rejected using the direction-specific dependence measure $I(\phi_R; \Delta_\tau \phi_C | \phi_C)$ indicating a significantly positive information flow from the respiration to the heart dynamics. Thus we can infer that the respiratory rhythm influences the heart rhythm.

9.7. Towards reliable inference of causality

In the construction of a test for causal relations, or, more precisely, for the direction of interaction from time series, the first step is a careful choice of the discriminating statistic. As we have seen in Sec. 9 9.3, not every asymmetric dependence measure is suitable. As noted by Paluš et al. [202], the direction of coupling can be inferred when two systems are coupled, but not yet fully synchronized. This can be understood considering the example of identical synchronization. Once the (identical) systems are synchronized, they produce identical time series and there is no way how to infer the correct causality relation just from the measured data. In the case of generalized synchronization, there is a one-to-one relation between the states of the systems. Time series $\{x(t)\}$ can be predicted from time series $\{y(t)\}$ and vice versa. Although some dependence measures, including those based on prediction errors, can give different values for the relations $x \to y$ and $y \to x$, these values are not given by the causality relations but rather by properties of the functional relation between the states of the systems, e.g. by its Jacobian. The causal relation can be inferred only when coupling is weaker than that necessary for emergence of synchronization, or when the synchronized state is frequently perturbed by variability in coupling or by internal or external noise driving the systems out of the synchronized state. Then the relation between the system states is not deterministic, but probabilistic, and can be measured by measures of statistical dependence, such as the above introduced information-theoretic measures.

Various asymmetric measures of dependence can have nonzero values even in the uncoupled direction in cases of unidirectional coupling. This holds for both probabilistic and deterministic measures. Even though no deterministic relation exists before the systems are synchronized, the deterministic measures, estimated in a coarse-grained approximation, reflect the statistical dependence which occurs in both direction even in the case of unidirectional coupling. Mutual comparison of these positive values or positivity/negativity of their difference (sometimes rescaled to some 'directionality indices') does not necessarily indicate the correct causal direction. For a correct inference of causality it is desirable to have a measure which vanishes in the uncoupled direction in the case of unidirectional coupling so that we can identify the causal direction by its statistically significant digression from zero, while in the uncoupled direction the measure does not cross the borders of a 'statistical zero'. The latter is given by the range obtained from appropriate surrogate data, separately for each direction. As a measure fulfilling this requirement we introduced the conditional mutual information (CMI).

The proper conditioning which assures the vanishing CMI in the uncoupled direction should contain full information about the future in the uncoupled state of the system, influence on which is evaluated. It means that in a case of m-dimensional dynamical system, or a variable which can be modelled by a (possibly nonlinear) autoregressive process of order m, the proper condition is an m-dimensional vector. The order m should be estimated from studied data before causality tests are applied. Then the estimation of $m + 2$ dimensional probability distribution functional can also be a nontrivial problem. It can be helpful, if the studied coupling can be reflected in the dynamics of instantaneous phases, since, in the case of phase dynamics, one-dimensional conditioning is sufficient in many cases.

Having an appropriate asymmetry dependence measure, asymptotically vanishing in the uncoupled direction, the inference of causality can be complicated by a bias in estimation from a limited amount of possibly noisy data. Therefore we need to establish a statistical significance of the obtained result, e.g., by the application of the surrogate data testing approach. The surrogate data, however, should reflect statistical and dynamical properties of the tested data, since those can be the source of bias. It is necessary to test that the surrogate data preserve the frequency distribution of the original data, which might be more important than the amplitude distribution. For instance, it might be entirely incorrect to make tests using the white noise (IID, scrambled) surrogates, obtained by the random permutation of amplitude time series, even though they preserve amplitude distributions. Exceptional care must be applied when

we study relations between systems which have different main frequencies, or different complexity of dynamics, or even different variability.

In many applications of the surrogate data testing, the significance of the departure of the tested values from the surrogate range is based on the (not always explicitly stated) assumption of a normal distribution of the test statistics estimated using the surrogate data. It is necessary, however, to study the surrogate distributions from large enough surrogate ensembles in order to establish the test significance independently of the form of the distribution of the discriminating statistics.

Before any real data application, it is always useful to assess the performance of any test, using appropriate model data, in order to estimate the amount of data necessary for reliable inference. Subsampled data can cause problems, while increasing the data amount by oversampling does not improve the test performance. As we demonstrated above, when evaluating the performance of a test, the Neyman-Pearson's hypothesis testing approach is appropriate. Having a possibility to perform a reliable test, secured by the appropriateness of both the discriminating statistic and the null hypothesis, inference of causality or any other phenomenon from a particular experimental time series of a sufficient length is based on the Fisher's significance testing approach. For processing a larger ensemble of experimental data, the Neyman-Pearson's hypothesis testing approach can also be appropriate. In such a task, the principal aim of the data processing is not an inference of a phenomenon or a property, but rather a distinction of groups according to a presence/absence of a property or a phenomenon. In any case, an appropriate null hypothesis should be chosen and carefully realized using suitable surrogate data. Interpreting the obtained results, provided one used an appropriate testing approach, it is necessary to keep in mind the used null hypothesis and its possible negations in order to consider physically plausible alternative hypotheses if the null hypothesis is improbable or rejected.

10. CONCLUSION

We have described the way how the statistical testing became a method frequently discussed in the physical literature related to nonlinear dynamics and the theory of deterministic chaos. The question 'Is it noise, or is it chaos?' asked about experimental data is a particular case of general questions of inference asked about processes underlying experimental data. The traditional field of mathematical statistics provides both a language and a toolbox for dealing with the questions of inference that emerge in analyses of data, recorded in complex systems for which the relation *theory – model – data* is not straightforward, but sophisticated statistical methods are used in order to distinguish repetitive patterns from random effects. In particular, we discussed the computationally intensive Monte Carlo approaches in which distributions of discriminating statistics are numerically obtained, using suitable models or randomization schemes. The latter provide the so-called surrogate data, numerical realizations of the considered null hypothesis.

Frequently in the physical literature, the surrogate data are understood as realizations of a linear stochastic process and serve as a testing ground for detecting nonlinear deterministic explanations of observed complex dynamics. We have discussed this concept, pointed out to many pitfalls that could lead to false detections of nonlinearity or deterministic chaos. We have proposed to test for specific nonlinear properties, really existing in the data, rather then to evaluate some abstract 'chaotic measures'. We have demonstrated our considerations in concrete examples in tests for nonlinearity in atmospheric data, sunspot numbers and brain signals. Further, we have generalized the notion of surrogate data as a general computational framework for numerical realizations of null hypotheses in inference of dynamical mechanisms and interactions from time series. We have described in detail an approach for detecting directional interactions, or causality from bivariate time series. Possible sources of false causality detections on various levels were discussed, from the choice of the discriminating statistics and suitable dynamical and/or statistical models, to the correct estimation of the test significance using the empirical distribution of the discriminating statistic under the null hypothesis. We have demonstrated that the correct application of the statistical testing in inference problems, emerging in the physics of complex systems, is far from trivial, however, can be of a non-negligible value when the tests are correctly performed and interpreted.

We hope that this review will be useful for physicists working in different fields and will inspire researchers to use the statistical tests as well as help to avoid underestimations of the results obtained by the statical inference from experimental data. The results inferred from experimental data should be considered as an inspiration for formulating new physical theories, rather then proposing theories without analysing data, since, as Sherlock Holmes said to Dr. Watson [231]: 'It is a capital mistake to theorize before one has data. Insensibly one begins to twist facts to suit theories, instead of theories to suit facts.'

Acknowledgements

The author would like to thank Aneta Stefanovska for the encouragement to write this article. A. Stefanovska, B. Musizza and the BRACCIA team (http://www.lancs.ac.uk/depts/physics/braccia/) are acknowledged for the cardiorespiratory data and the related cooperation. The author would also like to acknowledge the cooperation with D. Novotná, M. Vejmelka and K. Schindler-Hlaváčková.

This article reviewed a selection of results from several projects supported from different sources. In particular, the EC FP6 project BRACCIA (Contract No

517133 NEST), the Grant Agency of the Academy of Sciences of the Czech Republic projects No. IAA3042401 and IAA300420805, and Institutional Research Plans MSM002160849 and AV0Z10300504 are acknowledged.

[1] B. J. Mason, Contemp. Phys. **34 (1)** 19 (1993).
[2] B. J. Mason, Contemp. Phys. **43 (1)** 1 (2002).
[3] M. V. Mesquita, A. R. Vasconcellos and R. Luzzi, Contemp. Phys. **40 (4)** 247 (1999).
[4] M. Bier, Contemp. Phys. **46 (1)** 41 (2005).
[5] R. S. Conroy and C. Danilowicz, Contemp. Phys. **45 (4)** 277 (2004).
[6] S. A. Harris, Contemp. Phys. **45 (1)** 11 (2204).
[7] A. Stefanovska and M. Bracic, Contemp. Phys. **40 (1)** 31 (1999).
[8] C. N. Guy, Contemp. Phys. **37 (1)** 15 (1996).
[9] B. Ott, Contemp. Phys. **44 (1)** 1 (2003).
[10] J. D. Farmer, M. Shubik and E. Smith, Phys. Today **58 (9)** 37 (2005).
[11] M. W. Deem, Phys. Today **60 (1)** 42 (2007).
[12] G. Cowan, *Statistical Data Analysis*, (Oxford University Press, Oxford, 1998).
[13] S. N. Zhang and D. Ramsden, Experimental Astronomy **1 (3)** 145 (1990).
[14] L. Lyons and M. K. Unel, Eds., *Statistical Problems in Particle Physics, Astrophysics and Cosmology* (Imperial College Press, Covent Garden, London, UK, 2006)
[15] D. R. Tovey and N. J. T. Smith, J. Phys. G: Nucl. Part. Phys. **28** 2723 (2002).
[16] F. James, *Statistical Methods in Experimental Physics* (World Scientific, Singapore, 2006).
[17] J. T. Linnemann, arXiv:physics/0312059v2 (2003).
[18] K. Cranmer, arXiv:physics/0511028v2 (2006).
[19] N. Nicholls, Bulletin of the American Meteorological Society **81 (5)** 981 (2000).
[20] *Jozef Kvasnica 17 March 1930-23 November 1992*, Czechoslovak J. Phys. **43 (12)** 1145 (1993).
[21] M. Priestley, *Spectral Analysis and Time Series* (Academic Press, New York, 1981).
[22] M. B. Priestley, *Non-Linear and Non-Stationary Time Series Analysis* (Academic Press, London, 1988).
[23] H. Tong, *Non-Linear Time Series Analysis* (Oxford University Press, Oxford, 1990).
[24] H. Tong, Scandinavian Journal of Statistics, **22 (4)** 399 (1995).
[25] Shu-yu Zhang, *Bibliography on Chaos*, Directions in Chaos Vol. 5 (World Scientific, Singapore, 1991).
[26] `http://www.nytimes.com/books/98/`2/06/specials1 `/editorschoice87.html`
[27] J. Gleick, *Chaos. Making a New Science* (Viking, 1987).
[28] H. G. Schuster, *Deterministic Chaos. An Introduction.* (Physik Verlag, Weinheim, 1988).
[29] T. Leibler, Synthese **113** 357 (1998).
[30] J. Koperski, Brit. J. Phil. Sci. **52** 683 (2001).
[31] M. Peckham, *Man's Rage for Chaos: Biology, Behavior & the Arts* (Maisonneuve Press, Washington, 1997).
[32] R. T. Pascale, M. Millemann and L. Gioja, *Surfing the Edge of Chaos. The Laws of Nature and the New Laws of Business* (Crown Business, New York, 2000).
[33] M. Hénon, Commun. Math. Phys. **50** 69 (1976).
[34] N. H. Packard, J. P. Crutchfield, J. D. Farmer and R. S. Shaw, Phys. Rev. Lett. **45 (9)** 712 (1980).
[35] C. Nicolis and G. Nicolis, Nature **326** 523 (1987).
[36] K. Fraedrich, J. Atmos. Sci. **43** 419 (1986).
[37] C. Essex, T. Lockman and M. A. H. Nerenberg, Nature **326** 64 (1987).
[38] J. Kurths and H. Herzel, Physica D **25** 165 (1987).
[39] M.D. Mundt, W.B. Maguire II and R.R.P. Chase, J. Geophys. Res. A **96 (2)** 1705 (1991).
[40] M.N. Kremliovsky, Solar Phys. **151** 351 (1994).
[41] D. A. Hsieh, J. Finance **46** 1839 (1991).
[42] E. E. Peters, Financial Analysts J. **3** 55 (1991).
[43] G. DeCoster, W. Labys and D. Mitchell, J. Futures Markets **12** 291 (1992).
[44] A. L. Goldberger, D. R. Rigney, J. Mietus, E. M. Antman and S. Greenwald, Experientia **44** 983 (1988).
[45] D. R. Chialvo and J. Jalife, Nature **330** 749 (1987).
[46] A. Babloyantz and A. Destexhe, Proc. Nat. Acad. Sci. (USA) **83** 3513 (1986).
[47] P. E. Rapp, T. R. Bashore, J. M. Martinerie, A. M. ALbano and A. I. Mees, Brain Topogr. **2** 99 (1989).
[48] E. Başar (Editor), *Chaos in Brain Function* (Springer, Berlin, Heidelberg, 1990).
[49] F. Takens, In: D.A. Rand and D.S. Young (Editors), *Dynamical Systems and Turbulence (Warwick 1980)*, Lecture Notes in Mathematics 898. (Springer, Berlin, 1981), pp. 366–381.
[50] T. Sauer, J. Yorke and M. Casdagli, J. Stat. Phys. **65** 579 (1991).
[51] P. Grassberger and I. Procaccia, Physica D **9** 189 (1983).
[52] A. Wolf, J. B. Swift, H. L. Swinney and J. A. Vastano, Physica D **16** 285 (1985).
[53] M. Sano and Y. Sawada, Phys. Rev. Lett. **55** 1082 (1985).
[54] P. Bryant, R. Brown and H. D. I. Abarbanel, Phys. Rev. Lett. **65** 1523 (1990).
[55] R. Brown, P. Bryant and H. D. I. Abarbanel, Phys. Rev. A **43** 2787 (1991).
[56] G. Mayer-Kress (Editor), *Dimensions and Entropies in Chaotic Systems* (Springer, Berlin, 1986).
[57] M. Paluš, Neural Network World **7 (3)** 269 (1997). `http://www.cs.cas.cz/mp/papers/rd`a.pdf1
[58] P. Grassberger, T. Schreiber and C. Schaffrath, Int. J. Bifurcation and Chaos **1** 521 (1991).
[59] H. D. I. Abarbanel, R. Brown, J. J. Sidorowich and L. Sh. Tsimring, Rev. Mod. Phys. **65** 1331 (1993).
[60] D. Kugiumtzis, B. Lillekjendlie and N. Christophersen, Modeling, Identification and Control **15** 205 (1994).
[61] D. Kugiumtzis, B. Lillekjendlie and N. Christophersen, Modeling, Identification and Control **15** 225 (1994).
[62] H. D. I. Abarbanel, *Analysis of Observed Chaotic Data* (Springer, New York, 1996).
[63] H. Kantz and T. Schreiber, *Nonlinear Time Series Analysis* (Cambridge University Press, Cambridge, 1997).
[64] M. Casdagli and S. Eubank (Editors), *Nonlinear Modeling and Forecasting.* Santa Fe Institute Studies in the Science of Complexity, Proc. Vol. XII (Addison-Wesley, Reading, MA, 1992).

[65] A. S. Weigend and N. A. Gershenfeld (Editors), *Time Series Prediction: Forecasting the Future and Understanding the Past.* Santa Fe Institute Studies in the Science of Complexity, Proc. Vol. XV (Addison-Wesley, Reading, MA, 1993).

[66] H. Kantz, J. Kurths and G. Mayer-Kress (Editors), *Nonlinear Techniques In Physiological Time Series Analysis.* Springer Series in Synergetics, (Springer, Heidelberg, 1998).

[67] T. Schreiber, Phys. Rep. **308 (1)** 2 (1999).

[68] R. Hegger, H. Kantz and T. Schreiber, Chaos **9 (2)** 413 (1999)

[69] I. Dvorak and J. Siska, Phys. Lett. A **118 (2)** 63 (1986).

[70] J. Theiler, Phys. Rev. A **34 (3)** 2427 (1986).

[71] L. A. Smith, Phys. Lett. A **A 133 (6)** 283 (1988).

[72] A. Krakovská, Acta Physica Slovaca **45 (5)** 567 (1995).

[73] J. Theiler, Phys. Rev. A **41** 3038 (1990).

[74] R. L. Smith, J. R. Statist. Soc. B **54** 329 (1992).

[75] F. Takens, in B. L. J. Braaksma, H. W. Broer and F. Takens (Editors), *Dynamical Systems and Bifurcations*, Lecture Notes in Mathematics **1125** (Springer, Heidelberg, 1985).

[76] C. D. Cutler, J. Stat. Phys. **62** 651 (1991).

[77] C. D. Cutler, Philosoph. Trans. Royal Soc. London A **348** 343 (1995).

[78] A. R. Osborne and A. Provenzale, Physica D **35 (3)** 357 (1989).

[79] J. Theiler, Phys. Lett. A **155 (8-9)** 480 (1991).

[80] I. Dvorak and J. Klaschka, Phys. Lett. A **45 (5)** 225 (1990).

[81] A. Stefanovska, S. Strle and P. Kroselj, Phys. Lett. A **235 (1)** 24 (1997).

[82] J. Theiler and S. Eubank Chaos **3** 771 (1993).

[83] J. Theiler, S. Eubank, A. Longtin, B. Galdrikian and J.D. Farmer, Physica **D 58** 77 (1992).

[84] A. R. Osborne, A. D. Kirwin, A. Provenzale and L. Bergamosco, Physica D **23** 75 (1986).

[85] D. T. Kaplan and R. J. Cohen, Circullation Res. **67** 886 (1990).

[86] J. P. Pijn and F. H. Lopes da Silva, in I. Dvorak and A. Holden (Editors), *Mathematical Approaches to Brain Functioning Diagnostics.* Proceedings in Nonlinear Science. (Manchester University Press, Manchester and New York, 1991), pp. 337–352.

[87] A. K. Bera, in: C. R. Rao and G. J. Szekely (Editors), *Statistics for the 21th Century* (Marcel Dekker, New York, 2000), pp. 33–92.

[88] K. Pearson, Phil. Mag. Ser. 5 **50** 157 (1900).

[89] I. Hacking, Science **84 (5)** 67 (1984).

[90] R. P. Carver, Harvard Educational Review **48** 378 (1978).

[91] J. Cohen, American Psychologist **49** 997 (1994).

[92] D. H. Johnson, Journal of Wildlife Management **63 (3)** 763 (1999).

[93] R. A. Fisher, Nature **136** 474 (1935).

[94] R. A. Fisher, J. R. Stat. Soc. **17** 69 (1955). 69–78.

[95] R. Fisher, *Statistical Methods and Scientific Inference. 3rd Edn.* (Macmillan, New York, 1973).

[96] J. Neyman and E. Pearson, Phil. Trans. Roy. Soc. A **231** 289 (1933).

[97] J. O. Berger, Statistical Science **18 (1)** 1 (2003).

[98] R. S. Burington and D. C. May, *Handbook of Probability and Statistics with Tables.* (Handbook Publishers, Sandusky, Ohio, 1953).

[99] W. H. Beyer (Editor), *Handbook of Tables for Probability and Statistics.* (Blackwell, Oxford, 1966).

[100] P. J. Bickel and K. A. Doksum, *Mathematical Statistics* (Prentice Hall, New Jersey, 1977).

[101] G. A. Barnard, J. R. Statist. Soc. B **25** 294 (1963).

[102] A. C. A. Hope, J. R. Statist. Soc. B **30** 582 (1968).

[103] J. Besag and P. J. Diggle, Appl. Statist. **26** 327 (1977).

[104] P. Hall and D. M. Titterington, J. R. Statist. Soc. B **51** 459 (1989).

[105] E. W. Noreen, *Computer Intensive Methods for Testing Hypotheses.* (Wiley, New York, 1989).

[106] N. I. Fisher and P. Hall, Austral. J. Statist. **32** 177 (1990).

[107] R. S. Tsay, Appl. Statist. **41** 1 (1992).

[108] B. Efron, SIAM Review **21** 460 (1979).

[109] B. Efron and R. Tibshirani, Statist. Sci. **1** 54 (1986).

[110] D. V. Hinkley, J. R. Statist. Soc. B **50** 321 (1988).

[111] T. J. Diciccio and J. P. Romano, J. R. Statist. Soc. B **50** 338 (1988).

[112] B. Efron and R. Tibshirani, *An Introduction to the Bootstrap.* (Chapman and Hall, London, 1993).

[113] P. Hall and S. Wilson, Biometrics **47** 757 (1991).

[114] R. Beran, J. Am. Statist. Assoc. **83** 687 (1988).

[115] R. Wolf, Acad. Sci. Comp. Rend. **35** 704 (1852).

[116] D. Rind, Science **296** 673 (2002).

[117] R. P. Kane, Advances in Space Research **35(5)** 866 (2005).

[118] E. Bard and M. Frank, Earth and Planetary Science Letters **248(1-2)** 1 (2006).

[119] E. Friis-Christensen and K. Lassen, Science **254** 698 (1991).

[120] P.E. Damon and A. N. Peristykh, Geophys. Res. Lett. **26** 2469 (1999).

[121] P. Laut, J. Atmos. Solar-Terr. Phys. **65** 801 (2003).

[122] `http://www.realclimate.org/index.php/archives/2007/05/fun-with-correlations`

[123] `http://www.maths.manchester.ac.uk/~mrm/Teaching/OptoStats/slides/Why_Stats.pdf`

[124] `http://www.senate.gov/pagelayout/history/one_item_and_teasers/partydiv.htm`

[125] `ftp://ftp.ngdc.noaa.gov/STP/SOLAR_DATA/SUNSPOT_NUMBERS/YEARLY`

[126] J. A. Barnes, H. H. Sargent and P. V. Tryon, In: R. O. Pepin, J. A., Eddy and R. B. Merrill (Editors), *The Ancient Sun.* (Pergamon Press, New York, 1980), pp. 159-163.

[127] J. Theiler and D. Prichard, Physica **D 94** 221 (1996).

[128] M. Small and K. Judd, Physica D **120** 386 (1998).

[129] E. Mammen and S. Nandi, Phys. Rev. E **70** 016121 (2004).

[130] K. S. Chan, Fields Institute Commun **11** 77 (1997).

[131] J. Theiler, P. S. Linsay and D. M. Rubin, in [65], pp. 429–456.

[132] M. Paluš, Physica **D 80** 186 (1995).

[133] T. Schreiber and A. Schmitz, Phys. Rev. Lett. **77** 635 (1996).

[134] T. Schreiber and A. Schmitz, Physica **D 142(3-4)** 346 (2000).

[135] C.E. Shannon, Bell System Tech. J. **27** 379 (1948).

[136] T.M. Cover and J.A. Thomas, *Elements of Information Theory,* (J. Wiley, Sons, New York, 1991).

[137] K. Hlaváčková-Schindler, M. Paluš, M. Vejmelka and J. Bhattacharya, Phys. Rep. **441** 1 (2007).

[138] M. Paluš, Physica D **93** 64 (1996).

[139] Y. Horibe, IEEE Trans. on System, Man and Cybernetics **SMC-15**(5) 641 (1985).
[140] K. Fraedrich, J. Atmos. Sci. **43** 419 (1986).
[141] C. Essex, T. Lookman and M.A.H. Neremberg, Nature **326** 64 (1987).
[142] C.L. Keppenne and C.J. Nicolis, J. Atmos. Sci. **46** 2356 (1989).
[143] A.A. Tsonis and J.B. Elsner, Nature **333** 545 (1988).
[144] C. Nicolis and G. Nicolis, Nature **311** 529 (1984).
[145] P. Grassberger, Nature **323** 609 (1986).
[146] C. Nicolis and G. Nicolis, Nature **326** 523 (1987).
[147] P. Grassberger, Nature **326** 524 (1987).
[148] E. N. Lorenz, Nature **353** 241 (1991).
[149] M. Paluš and D. Novotná, Phys. Lett. A **193** 67 (1994).
[150] M. Paluš, V. Albrecht and I. Dvořák, Phys. Lett. A **175** (1993) 203.
[151] M. Paluš, in: [65], pp. 387–413.
[152] R.G. Miller, *Simultaneous Statistical Inference* (Springer New York-Heidelberg-Berlin, 1980).
[153] F. Mosteller and J.W. Tukey, *Data Analysis and Regression* (Addison-Wesley, Reading, Mass., 1977).
[154] G. Hommel, Biometrical J. **25** (1983) 423.
[155] D. Morgerstern, Metrika **27** (1980) 171.
[156] D. Prichard and J. Theiler, Physica D **84** 476 (1995).
[157] G. Sugihara, M. Casdagli, E. Habjan, D. Hess, P. Dixon and G. Holland, PNAS **96 (25)** 14210 (1999).
[158] M. Paluš, D. Novotná and P. Tichavský, Geophys. Res. Lett. **32** L12805 (2005).
[159] J. W. Hurrell, Y. Kushnir and M. Visbeck, Science **291(5504)** 603 (2001).
[160] D. J. Thomson, Science **268** 59 (1995).
[161] K.L. Harvey (Editor), *The Solar Cycle,* (Astronomical Society of the Pacific Conference Series 27, San Francisco, 1992).
[162] F. Krause and K.-H. Rädler, *Mean-Field Magnetohydrodynamics and Dynamo Theory* (Pergamon Press, Oxford, 1980).
[163] M.R.E. Proctor and A.D. Gilbert (Editors), *Lectures on Solar and Planetary Dynamos* Cambridge University Press, 1994.
[164] M. Ossendrijver, Astronomy and Astrophysics Review **11(4)** 287 (2003).
[165] U. Feudel, W. Jansen and J. Kurths, Int. J. Bif. & Chaos **3** 131 (1993).
[166] M. Küker, R. Arlt and G. Rüdiger, Astronomy and Astrophysics **343** 977 (1999).
[167] N.O. Weiss and S.M. Tobias, Space Science Reviews **94** 99 (2000).
[168] J. R. Kuhn, Advances in Space Research **34** 302 (2004).
[169] G.L. Withbroe, Spacecraft **26** 394 (1989).
[170] M. Dämmig and F. Mitschke, Phys. Lett. A **178** 385 (1993).
[171] M. Paluš, in: H. Kantz, J. Kurths and G Mayer-Kress (Editors), *Nonlinear Analysis of Physiological Data.* (Springer Series in Synergetics, Heidelberg, 1998).
[172] SIDC-team, Royal Observatory of Belgium, Ringlaan 3, B-1180 Brussels, Belgium. Internet address: `http://sidc.oma.be/DATA/yearssn.dat`
[173] N.O. Weiss, Philos. Trans. R. Soc. London, Ser. A **330** 617 (1990).
[174] L.A. Smith, In: G. Cini Castagnoli and A. Provenzale (Editors), *Past and Present Variability of the Solar-Terrestrial System: Measurement, Data Analysis and Theoretical Models* (IOS Press Ohmsha, Amsterdam, 1997), pp. 177–246.
[175] D. Gabor, J. IEE London **93** 429 (1946).
[176] M. G. Rosenblum, A. S. Pikovsky and J. Kurths, Phys. Rev. Lett. **76** 1804 (1996).
[177] M. Waldmeier, Astron. Mitt. Zürich **14(133)** 105 (1935).
M. Waldmeier, Astron. Mitt. Zürich **14(138)** 470 (1939).
[178] D.H. Hathaway, R.M. Wilson and E.J. Reichman, Solar Physics **151** 177 (1994).
[179] W.H. Press, B.P. Flannery, S.A. Teukolsky and W.T. Vetterling, *Numerical Recipes: The Art of Scientific Computing* (Cambridge Univ. Press, Cambridge, 1986).
[180] SIDC-team, Royal Observatory of Belgium, Ringlaan 3, B-1180 Brussels, Belgium. Internet address: `http://sidc.oma.be/DATA/monthssn.dat`
[181] M. Paluš, J. Kurths, U. Schwarz, N. Seehafer, D. Novotná and I. Charvátová, Phys. Lett. A **365** 421 (2007).
[182] M. Paluš and D. Novotná, Phys. Rev. Lett. **83(17)** 3406 (1999).
[183] J. Timmer, Phys. Rev. Lett. **85 (12)** 2647 (2000).
[184] M. Paluš and D. Novotná, Phys. Rev. Lett. **85 (12)** 2648 (2000).
[185] P. D. Mininni, D. O. Gomez and G. B. Mindlin, Solar. Phys. **208** 167 (2002).
[186] B. H. Jansen, Int. J. Biomed. Comput. **27** 95 (1991).
[187] W. J. Freeman, Int. J. Bifurcations and Chaos **2** 451 (1992).
[188] C. J. Stam, Clin. Neurophysiol. **116(10)** 2266 (2005).
[189] J. Theiler and P. E. Rapp, EEG Clin. Neurophysiol. **98** 213 (1996).
[190] W. S. Pritchard, D. W. Duke, and K. K. Krieble, Psychophysiology **32** 486 (1995).
[191] M. Paluš, Biol. Cybern. **75** 389 (1996).
[192] D. R. Chialvo and M. M. Millonas, Phys. Lett. A **209** 26 (1995).
[193] M. Paluš, I. Dvořák and I. David, Physica A **185** 433 (1992).
[194] A. Pikovsky, M. Rosenblum and J. Kurths, *Synchronization. A Universal Concept in Nonlinear Sciences* (Cambridge University Press, Cambridge, 2001).
[195] C. Schäfer, M.G. Rosenblum, J. Kurths and H.-H. Abel, Nature **392** 239 (1998); Phys. Rev. E **60** 857 (1999).
[196] M. Paluš, D. Hoyer, IEEE Engineering in Medicine and Biology **17(6)** 40 (1998).
[197] M. Bračič Lotrič and A. Stefanovska, Physica A **283(3-4)** 451 (2000).
[198] A. Stefanovska, H. Haken, P. V. E. McClintock, M. Hožič, F. Bajrović and S. Ribarič, Phys. Rev. Lett. **85(22)** 4831 (2000).
[199] S.J. Schiff, P. So, T. Chang, R.E. Burke and T. Sauer, Phys. Rev. E **54**, 6708 (1996).
[200] M. Le Van Quyen, J. Martinerie, C. Adam and F.J. Varela, Physica D **127**, 250 (1999).
[201] P. Tass, M.G. Rosenblum, J. Weule, J. Kurths, A. Pikovsky, J. Volkmann, A. Schnitzler and H.-J. Freund, Phys. Rev. Lett. **81** 3291 (1998).
[202] M. Paluš, V. Komárek, Z. Hrnčíř and K. Štěrbová, Phys. Rev. E **63** 046211 (2001).
[203] M. Paluš, V. Komárek, T. Procházka, Z. Hrnčíř and K. Štěrbová, IEEE Engineering in Medicine and Biology Magazine **20(5)** 65 (2001)
[204] N. Wiener, in: E. F. Beckenbach (Editor), *Modern*

Mathematics for Engineers (McGraw-Hill, New York, 1956).
[205] C.W.J. Granger, Econometrica **37** 424 (1969).
[206] J. Geweke, in: Z. Griliches and M.D. Intriligator (Editors), *Handbook of Econometrics* (North-Holland, 1984) vol. 2, 1101–1144.
[207] U. Triacca, Theor. Appl. Climatol. **81** 133 (2005).
[208] M. Kamiński, M. Ding, W.A. Truccolo and S.L. Bressler, Biol. Cybern. **85** 145 (2001).
[209] K.J. Blinowska, R. Kuś and M. Kamiński, Phys.Rev. E **70** 050902(R) (2004)..
[210] J. Arnhold, P. Grassberger, K. Lehnertz and C.E. Elger, Physica D **134** 419 (1999).
[211] T. Schreiber, Phys. Rev. Lett. **85** 461 (2000).
[212] R. Quian Quiroga, J. Arnhold and P. Grassberger, Phys. Rev. E **61(5)** 5142 (2000).
[213] M.G. Rosenblum and A.S. Pikovsky, Phys. Rev. E **64** 045202(R) (2001).
[214] M.G. Rosenblum, L. Cimponeriu, A. Bezerianos, A. Patzak and R. Mrowka, Phys. Rev. E **65** 041909 (2002).
[215] M. Paluš and A. Stefanovska, Phys. Rev. E **67** 055201R (2003).
[216] K. Otsuka, Y. Miyasaka and T. Kubota *et al.*, Phys. Rev. E **69(4)** 046201 (2004).
[217] P.F. Verdes, Phys. Rev. E **72(2)** 026222 (2005).
[218] I.I. Mokhov and D.A. Smirnov, Geophys. Rev. Lett. **33(3)** L03708 (2006)
[219] R. Quian Quiroga, A. Kraskov, T. Kreuz and P. Grassberger, Phys. Rev. E **65** 041903 (2002).
[220] M. Chávez, J. Martinerie and M. Le Van Quyen, Journal of Neuroscience Methods **124** 113 (2003).
[221] J. Brea, D.F. Russell and A.B. Neiman, Chaos **16(2)** 026111 (2006).
[222] R. Marschinski and H. Kantz, Eur. J. Phys. B **30(2)** 275 (2002).
[223] M. Paluš and M. Vejmelka, Phys. Rev. E **75** 056211 (2007).
[224] K. Pyragas, Phys. Rev. E **56** 5183 (1996).
[225] N. F. Rulkov, M. M. Sushchik, L. S. Tsimring and H. D. I. Abarbanel, Phys. Rev. E **51**, 980 (1995).
[226] R. Quian Quiroga's FORTRAN code from `http://www.vis.caltech.edu/~rodri/programs/synchro.for`
[227] M. Paluš, Phys. Lett. A **235**, 341 (1997).
[228] D. Prichard and J. Theiler, Phys. Rev. Lett. **73** 951 (1994).
[229] M. Paluš, Phys. Lett. A **213** 138 (1996).
[230] B. Musizza, A. Stefanovska, P. V. E. McClintock, et al., J. Physiol. (London) **580 (1)** 315 (2007).
[231] Sir A. Conan Doyle, *The Scandal in Bohemia. In: The Adventures of Sherlock Holmes* (George Newnes Ltd., London, 1892).
[232] Obviously, for a linear function F, the model (13) is a special case of (10).

Milan Paluš is a senior researcher and a leader of the nonlinear dynamics group with the Institute of Computer Science of the Academy of Sciences of the Czech Republic in Prague. He was born in 1963 in Bojnice, Slovakia. In 1986 he finished his study of mathematical physics at the Charles University in Prague, in 1992 he received his CSc (PhD equivalent) degree in computer science from the Czechoslovak Academy of Sciences. In 1992–1994 he was a postdoc with the Santa Fe Institute. His main research interests are nonlinear dynamics, deterministic chaos, information theory and synchronization of complex systems, oriented namely to analysis of experimental time series with applications in diverse fields from physics, geophysics, meteorology, climatology, air pollution, through electrophysiology, neurology, to finance.