Transformation-cost time-series method for analyzing irregularly sampled data

Ibrahim Ozken,[1,2,*] Deniz Eroglu,[2,3,†] Thomas Stemler,[4] Norbert Marwan,[2] G. Baris Bagci,[5] and Jürgen Kurths[2,3,6]

[1]*Department of Physics, Ege University, 35100 Izmir, Turkey*
[2]*Potsdam Institute for Climate Impact Research (PIK), 14473 Potsdam, Germany*
[3]*Department of Physics, Humboldt University, 12489 Berlin, Germany*
[4]*School of Mathematics and Statistics, The University of Western Australia, 35 Stirling Highway, Crawley, WA 6009, Australia*
[5]*Department of Materials Science and Nanotechnology Engineering, TOBB University of Economics and Technology, 06560 Ankara, Turkey*
[6]*Institute for Complex Systems and Mathematical Biology, University of Aberdeen, Aberdeen AB24 3UE, United Kingdom*

(Received 20 April 2015; published 18 June 2015)

Irregular sampling of data sets is one of the challenges often encountered in time-series analysis, since traditional methods cannot be applied and the frequently used interpolation approach can corrupt the data and bias the subsequence analysis. Here we present the TrAnsformation-Cost Time-Series (TACTS) method, which allows us to analyze irregularly sampled data sets without degenerating the quality of the data set. Instead of using interpolation we consider time-series segments and determine how close they are to each other by determining the cost needed to transform one segment into the following one. Using a limited set of operations—with associated costs—to transform the time series segments, we determine a new time series, that is our transformation-cost time series. This cost time series is regularly sampled and can be analyzed using standard methods. While our main interest is the analysis of paleoclimate data, we develop our method using numerical examples like the logistic map and the Rössler oscillator. The numerical data allows us to test the stability of our method against noise and for different irregular samplings. In addition we provide guidance on how to choose the associated costs based on the time series at hand. The usefulness of the TACTS method is demonstrated using speleothem data from the Secret Cave in Borneo that is a good proxy for paleoclimatic variability in the monsoon activity around the maritime continent.

DOI: 10.1103/PhysRevE.91.062911

PACS number(s): 05.45.Tp, 05.10.−a, 92.60.Iv

I. INTRODUCTION

One of the challenges in time-series analysis is to detect dynamical changes in the evolution of the underlying system. We consider a dynamical system $\dot{\vec{x}} = f(\vec{x},p)$ with $\vec{x} \in \mathbb{R}^m$ and p a control parameter that depends on time $p = g(t)$. We want to detect in the scalar time series $s(t) = [s_0, \dots, s_N] = [M(x_0), \dots, M(x_N)]$, $M : \vec{x} \in \mathbb{R}^m \to s \in \mathbb{R}$ dynamical regime changes caused by the time dependence of p.

There are numerous methods that can be used to detect such regime changes in regularly sampled times series, i.e., the time resolution $\Delta t = t(s_{i+1}) - t(s_i) = \text{const}\ \forall i \in [0, N-1]$; see, e.g., Refs. [1–3]. However, in several disciplines, such as astrophysics and earth sciences, a constant sampling cannot be ensured. Therefore, regularly interpolation as a preprocessing step is often applied, but this might lead to a bias of the results [4,5]. For example, interpolation leads to a positive bias in autocorrelation estimation (and, thus, an overestimation of the persistence time) and a negative bias in cross correlation analysis [4].

Here we are exploring a different approach that can be used for irregularly sampled data without interpolation. Focussing on the paleoclimate time series, we are in particular interested in regime changes. Such paleoclimate proxy records show a very erratic sampling, with the sampling times often γ distributed [4,5], and in addition the records are subject to measurement noise, which make interpolation difficult. Instead of interpolating the time series we determine TACTS between segments of the original time series, which results in a new transformation-cost time series having regular sampling. This transformation-cost time series can then be analyzed by established methods. Wanting to detect regime changes we apply recurrence plot analysis being one of the appropriate methods for this purpose [6].

Our new approach is based on a measure introduced by Victor and Purpura [7] and which was further developed in order to transform spike trains to real-valued time series with regular resolution by Hirata and Aihara [8]. The idea behind this approach is similar to the FLUS method, which is: if the time series is from one dynamical regime, the cost of transformation from one segment to the subsequent one should be similar for each segment of the data [3]. Since, for example, in paleoclimate proxy records, we may not have any knowledge about the current control parameters, we compute the cost of transforming one segment into the following one. Dramatic changes in the cost time series indicate a change in the underlying dynamics.

While recurrence plot-based quantification is not directly applicable to irregularly sampled time series, we show that identifying regime changes in the dynamics of the system becomes possible by combining the TACTS approach with recurrence analysis.

Our paper is organized as follows: In the following section we introduce the technique of our time-series analysis and highlight how the parameters of the cost-transformation function can be determined based on measurement data. In Sec. III we apply our method using numerical data from paradigmatic model systems: the logistic map and the Rössler system. We evaluate the performance of our method using irregular sampling and measurement noise. As an application

[*] ibrahim.ozken@ege.edu.tr
[†] eroglu@pik-potsdam.de

we analyze a paleoclimate record in Sec. IV, which is a speleothem record from Borneo, Indonesia, representing the variability of the Indonesian-Australian monsoon over the past 62,000 years. Finally, we state a conclusion.

II. TIME-SERIES ANALYSIS

A. Metric analyses

In time-series analyses, an important issue is to calculate the distance between two data patterns. This distance problem frequently occurs, for example, in recurrence plots (RP) [6], the estimation of the maximum Lyapunov exponent [9], scale-dependent correlations [10], data classification [11], or correlation dimension estimations [12]. When doing these kinds of analyses and the data has an equidistant time resolution, the Euclidean distance is often used. However, in applications that generate data with nonregular sampling, such an approach is not directly applicable. Such data sets include almost all paleoclimate observations, which can have a very erratic almost random time resolution. One way to deal with such data sets is to interpolate them. Such an interpolation will not only fill the gaps but replace real measurements with new interpolated data points close by that have regular sampling. But this is often not the optimum method, since the subsequent analysis will be typically biased [13]. Moreover, these interpolated values have a higher uncertainty than the measured data points they replace.

For the dynamics of firing neurons, Victor and Purpura [7] showed that the spike time distance is a useful method that applies to irregularly sampled data sets. The basic idea of this method is a distance metric that provides information of how easily one data segment can be transformed into another one. To transform one segment of data into another segment, three elementary operations are required: adding or deleting of a data point and moving the data point to a different time. Using associated costs for these elementary operations, an optimal data transformation will be achieved if the cost of the transformation is minimized. We will illustrate this method for spike train data below before introducing our modified method for continuous data that determines TACTS.

Consider the metric D as a mapping of two pairs of spike trains or data segments, say S_a and S_b, onto a real value. In order for D to be metric, it must satisfy the following three conditions since D is a generalized distance:

(1) $D(S_a,S_b) \geqslant 0$ (positive)

(2) $D(S_a,S_b) = D(S_b,S_a)$ (symmetric)

(3) $D(S_a,S_c) \leqslant D(S_a,S_b) + D(S_b,S_c)$ (triangle inequality)[7]

In Fig. 1 we give one illustrative example how the elementary operations transform the spike train S_a into S_b [7]. When we transform S_a to S_b, the total cost is the sum of the elementary step's costs. As we can see, the transformation of S_a to S_b requires 7 distinct steps. Steps 1, 2, 4, 5, and 7 move one spike to a different time point, while step 3 deletes one spike, and in step 6 one spike is created. Assigning costs to each of these operations and pairwise checking the segments, Victor and Purpura [7] analyzed different types of spike time series. They chose the cost of deleting and adding to be the same $p_d = p_a = 1$, while moving a spike is proportional to the time distance that the data point is moved from t_a to t_b: $\lambda_0|t_a - t_b|$. Clearly the parameter λ_0 is a frequency with the unit Hz.

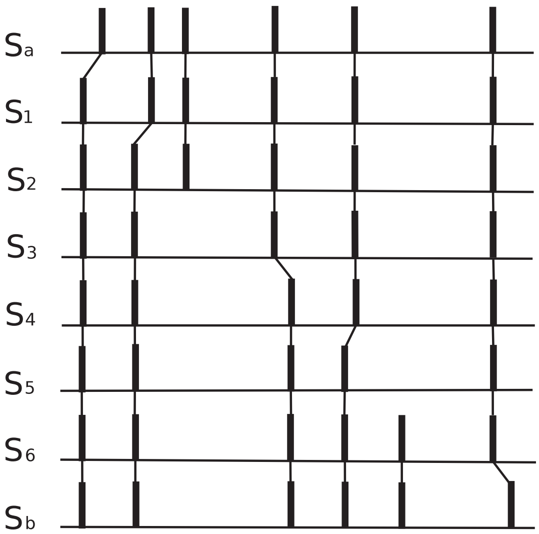

FIG. 1. Illustration of the transformation of S_a to S_b. In total S_a undergoes seven steps numbered $S_1, S_2, \ldots, S_6$. Note that $S_7 = S_b$. The path shown is a minimal-cost path and all the steps are elementary steps, like moving a spike or deleting and creating.

So far we only considered spike trains as they are apparent in the analysis of brain dynamics data. Suzuki *et al.* [14] have extended this method for continuous marked data and we follow their approach. The continuous data set is transformed to an event time series that is very similar to spike trains and allows events to have different amplitudes. The transformation from continuous data to an event time series is done by the system itself that is event occurrence as earthquakes, data quality as paleoclimate time series, extreme events as crisis in finance, etc. We define the cost function in the following way:

$$\begin{aligned} p(c) = &\sum_{(\alpha,\beta)\in C} \{\lambda_0|t_a(\alpha) - t_b(\beta)| \\ &+ \frac{1}{m}\sum_{k=1}^{m} \lambda_k|L_{a,k}(\alpha) - L_{b,k}(\beta)|\} \\ &+ \lambda_S(|I| + |J| - 2|C|), \end{aligned} \tag{1}$$

where I and J are a set of indices of the events in S_a and S_b, respectively. Note that the summation is over the pairs $(\alpha,\beta) \in C$, where C is the set of points that will be shifted in time. α and β are the αth event in S_a and the βth event in S_b. The first term with the coefficient λ_0 is the cost to shift some event in time [14]. The second summation involves the difference $|L_{a,k}(m) - L_{b,k}(n)|$, where $L_{a,k}$ is the amplitude of the kth event in S_a. Therefore, the parameter λ_k has the unit of amplitude $^{-1}$ and the sum is over the different components of the amplitude. That is, if we are dealing with one-dimensional data $m = 1$, while for a three-dimensional phase space $m = 3$. The last terms in the cost function deal with the events not in C which have to be added or deleted. Note that $|\cdot|$ denotes the cardinality of the set and λ_S is the cost parameter for this operation. Suzuki *et al.* omitted this parameter, since they chose a cost of one for such an operation [14].

As we can see minimization of the cost function, Eq. (1), will depend on the choice of $\lambda_{0,k,S}$. In Refs. [7] and [14] it was shown that $\lambda_{0,k}$ can be optimized to get excellent results in many cases. Here we are more concerned with time series that (a) were produced for a whole control parameter range (see Secs. III A and III B) or (b) are nonstationary (see Sec. IV). Instead of trying to optimize $\lambda_{0,k}$ for each time series, we suggest two possible new ways to choose these values and explore one of them in detail.

Note that in Eq. (1) the cost to delete or add a new data point is equal to λ_S for each of both operations. On the other hand, shifting and changing the amplitude is proportional to the difference in time and amplitude. In order to determine whether the deleting or adding operations are preferred, a natural choice for $\lambda_{0,k}$ should be such that every time shifting and changing cost more than $2\lambda_S$ (total cost of adding and deleting a point). Consequently $\lambda_{0,k}$ should be chosen in such a way that in average both terms give a contribution of λ_S. Therefore:

$$\lambda_0 = \frac{M}{\text{total time}}, \tag{2a}$$

$$\lambda_k = \frac{M-1}{\sum_i^{M-1} |L_{a,k} - L_{b,k}|}, \tag{2b}$$

where M is the total number of events in the time series. Hence, our first possibility chooses λ_0 as the mean event frequency and λ_k as the inverse of the average amplitude difference.

The second option we are focusing on also uses $\lambda_{0,k}$ according to Eqs. (2) but optimizes the cost of the deleting and adding operations. That is, while $\lambda_{0,k}$ are fixed, we explore the costs for deleting and adding in the range from $\lambda_S \in [0,4]$. If our time series consist of $n+1$ segments of equal length, we therefore can calculate n costs for each individual transformation of the segments. Assuming that the costs are linearly independent, the central limit theorem indicates that the costs should be normally distributed. Especially when dealing with nonstationary data, changing λ_S such that the distribution becomes normal strongly improves the skill of our time-series analysis method.

Following the outlined method we are able to determine TACTS for irregularly sampled data. First we divide a time series into equidistant segments and then calculate the costs between each consequent ones. Therefore, TACTS is created for whole time series and using a constant length of the data segments leads to a constant sampling. This new sampling rate is the length of the data segment and the value is determined by Eq. (1). Note that this implies another optimization since a large segment size will lead to a transformation time series of short length, while a short segment size might make it difficult to detect regime transitions.

Assume we have a time series: $X = (x_{t_1}, x_{t_2}, x_{t_3}, \ldots\ldots, x_{t_N})$, where N is the number of points and $t_2 - t_1 \neq t_3 - t_2 \neq \ldots \neq t_N - t_{N-1}$. We divide the time series to a set of segments W, which have equal size. After this, we have n equal windows $W_1, W_2, \ldots, W_n$ and we determine the transformation costs $p(W_1, W_2)$, $p(W_2, W_3)$,..., $p(W_{n-1}, W_n)$ for all sequence windows. This leads to a new equidistant time series, which is our transformation-cost time series. By using RP we detect regime changes in the underlying dynamics.

An intuitive understanding of the transformation-cost time-series method is based on an interpretation of the cost function Eq. (1) and the cost coefficients $\lambda_{0,k,S}$ Eq. (2). As mentioned above, $\lambda_{0,k}$ are the average amplitude and the average event frequency while λ_S penalizes changes in time and amplitude of an event that are large. These coefficients weigh the local difference between the event pairs in our cost function Eq. (1). Therefore, we can perceive the cost function as balancing the time and amplitude differences of the events in the two segments (S_a and S_b in Fig. 1) versus deleting and re-creating all events. If we analyze several segments resulting from a regular dynamics, the local difference between the segments will be bounded and the cost time series will show some regularity. If the underlying dynamics on the other hand is erratic, the local difference between the segments can be large and consequently the cost function shows no obvious regularity (this property is also used for the FLUS method [3]). Differencing a (regular) time series (applied, e.g., as a high-pass filter), $x_t - x_{t-1}$, is a special form of this approach. To measure regularity in our transformation-cost time series, we apply recurrence quantification analysis.

B. Recurrence plot

Recurrence plot was first introduced by Eckman *et al.* as a tool to visualize the recurrences of dynamical systems [15]. Assume we have an m-dimensional system, a state in this m-dimensional state space is ε-recurrent if its state vector falls for a certain $\varepsilon > 0$ into the neighborhood of another state vector. For a given trajectory $\vec{x}_i$ ($i = 1, \ldots, N, \vec{x}_i \in \mathbb{R}^m$), the recurrence plot **R** is defined as

$$R_{i,j}(\varepsilon) = \Theta(\varepsilon - \|\vec{x}_i - \vec{x}_j\|),\ i,j = 1, \ldots, N, \tag{3}$$

where $\Theta(\cdot)$ is the Heaviside function, and $\|\cdot\|$ is a norm [6]. Therefore, $R_{i,j} \equiv 1$ if the states at times i and j are recurrent, and $R_{i,j} \equiv 0$ otherwise.

Clearly, on the main diagonal of the RP $R_{i,i} \equiv 1$, which therefore is called the line of identity (LOI). RP matrices are symmetric (if norm is used for calculating the distance between states to create RP), binary matrices. Off-diagonal structures that are parallel to the LOI appear as line segments. These structures represent typical dynamical properties. For white noise we observe homogeneously distributed single recurrence points, while for deterministic dynamics diagonal line segments (parallel to the LOI) will dominate. The distribution of the line length can be used to distinguish between different dynamical regimes. Chaotic dynamics causes mainly rather short line segments and regular (periodic) dynamics causes very long line segments [2,6].

In order to study the dynamical features of different systems using the relationship between the system's dynamics and the distribution of line segments, several complexity measures based on line segments have been introduced as recurrence quantification analysis (RQA) [2,6]. The frequency distribution of diagonal line lengths $P(\ell)$ is directly linked with the dynamics, hence related with the Lyapunov exponent, since $P(\ell)$ quantifies the divergence behavior of the dynamical system.

One of the most important measures of RQA is the determinism (DET), which is quantifying the fraction of recurrence points $R_{i,j} \equiv 1$, which form diagonal lines [6],

$$\mathrm{DET} = \frac{\sum_{\ell=\ell_{\min}}^{N} \ell P(\ell)}{\sum_{\ell=1}^{N} \ell P(\ell)}. \tag{4}$$

DET is a good measure to detect periodic-chaotic regime transitions, since the measure based on $P(\ell)$ is related to predictability. Given our main motivation to find such regime transitions in our transformation-cost time series, we focus only on determinism, although other measures are appropriate as well [6].

III. NUMERICAL EXAMPLES

In real-world applications, especially in the paleoclimate data sets, time series are not equidistantly sampled. In order to deal with this kind of difficulty, we have created prototypical irregularly sampled models. Moreover, for the high possibility of noise in real-world applications, we have added noise into our study on the logistic map.

A. Logistic map

As our first application we analyze data from the logistic map, which is defined as

$$x_{i+1} = r x_i (1 - x_i) \tag{5}$$

for $r \in [0,4]$.

It has been shown [2,16–19] that analyzing RPs is an efficient method to detect the regime transitions in the logistic map's dynamics. We are going to analyze the dynamics and its transitions in a control parameter range of $r \in [3.5,4]$. For our investigation we sample the control parameter range with a step size of 0.001 and calculate a time series of 3000 iterations for each control parameter value. We delete the first 1000 points to discard transients, resulting in a time series consisting of 2000 points that have been used for all analysis of the logistic map in this paper.

We investigate the performance of our method for nonequidistant sampled data by deleting randomly 100 ($\gamma = 5\%$), 200 ($\gamma = 10\%$), 300 ($\gamma = 15\%$), or 400 points ($\gamma = 20\%$) from the original time series. For all time series we choose a segment size of four time steps. This size can still capture changes in the underlying dynamics even for $\gamma = 20\%$ but also results in a long enough transformation-cost time series that can be analyzed using RP. We determine the transformation cost for each window pair in the data set using $\lambda_{0,k}$ given from Eqs. (2) and optimized λ_S as outlined in Sec. II A. The value of λ_S depends on the particular γ and does decrease with increasing γ. While one could argue that for each chaotic regime one should use a different λ_S, we chose to determine only one value from the time series generated at $r = 4$. There are two reasons for this particular choice: (i) using one λ_S for all time series resembles the situation when no additional information about the control parameter is available and (ii) it shows that our method does not crucially depend on the choice but is stable even if λ_S is close enough to the optimum value. For our different γ levels we determined λ_S to be: $\lambda_S = 1.07$ ($\gamma = 5\%$), 1.04 (10%), 0.95 (15%), and 0.93 (20%).

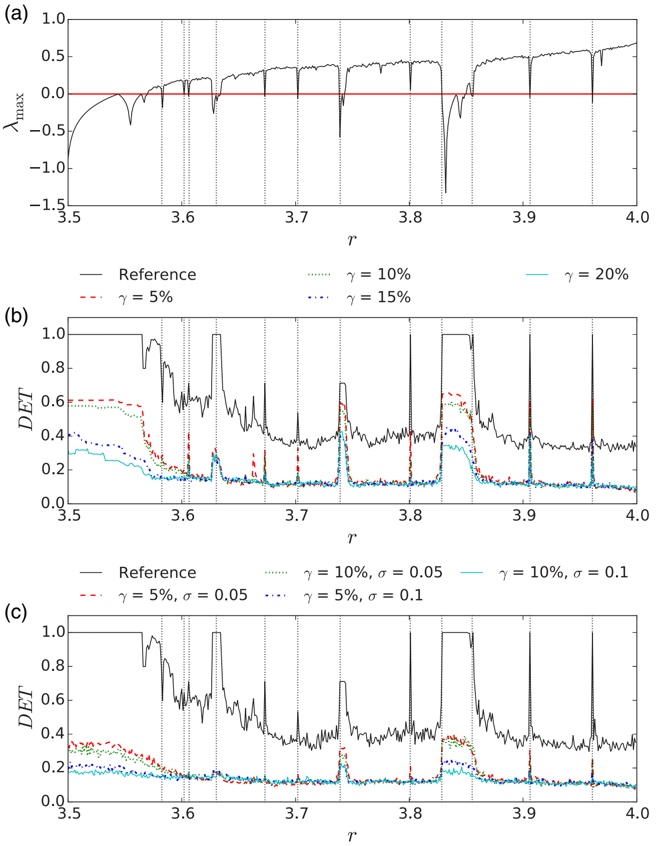

FIG. 2. (Color online) RQA analyses for logistic map: (a) Lyapunov exponent of the logistic map; (b) and (c) determinism calculated from TACTS for (b) various levels of deleting and (c) two measurement noise levels ($\sigma = 0.05$ and 0.1) and for two different rates of irregularity ($\gamma = 5\%$ and 10%). For details, see text as well as the legends of the figures.

This optimization results in $\lambda_S \approx 1.00$ and therefore is similar to the original values used in Refs. [7] and [14]. This transformation-cost time series is then used in the RP to calculate the determinism Eq. (4) with $\varepsilon = 0.08$ for all r values considered. The ε value needs to be sufficiently small [6] and $\varepsilon = 0.08$ is adequately suitable for our phase space of the transformation cost.

In Fig. 2 we present our results. Figure 2(a) shows the Lyapunov exponent calculated from the time series. In Fig. 2(b) the determinism calculated for the time series is shown for increasingly irregular sampling. Comparison with the Lyapunov exponent shows that the determinism tracks the transitions of the dynamics for all data sets. All abrupt drops from positive Lyapunov exponent to negative ones are clearly shown in the determinism measure cases. These drops are demonstrated with dotted lines in Fig. 2. We clearly see that randomly deleting points leads to a distinct drop in the determinism. While this drop is most pronounced when comparing the reference line ($\gamma = 0\%$) with $\gamma = 5\%$ for higher γ values the determinism does not decrease as much. This is due to the fact that any deleting will disconnect the long diagonal lines that contribute mainly to Eq. (4). Naturally the results are most conclusive for $\gamma = 0\%$, but even for $\gamma = 20\%$

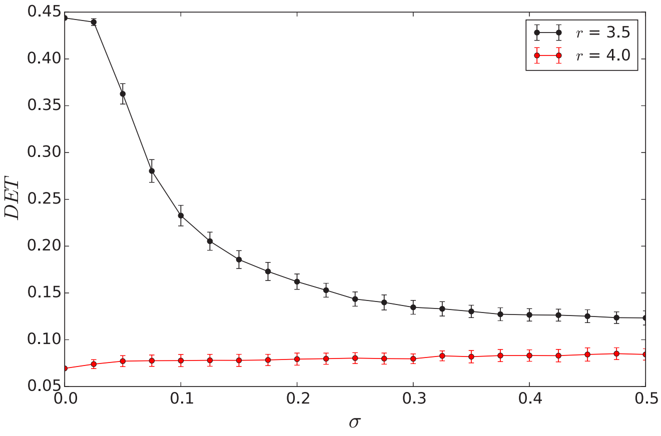

FIG. 3. (Color online) Determinism against noise level σ: $\gamma = 10\%$ and $r = 3.5$ (periodic, black line) and $r = 4$ (chaotic, red line).

we are able to identify the bifurcations in the logistic map and successfully detect the changes in the dynamics.

In Fig. 2(c), we show that our method cannot only detect changes in the dynamics for irregular sampling but is also stable if the data is additionally compromised by measurement noise. We added Gaussian white noise [$\langle\xi\rangle = 0$ and $\langle\xi(t)\xi(t')\rangle = \delta(t-t')$] and the results shown are for a noise level of $\sigma = 0{,}0.05$ and 0.1 ($\gamma = 5\%$ and 10%). Note that the σ is scaled relative to the variance of the time series. Again, we clearly see that for all noise levels considered, our method is able to identify the changes in the dynamics and closely follows the Lyapunov exponent.

To further investigate the stability of our method we investigate two time series of 2000 points corresponding to $r = 3.5$ (periodic dynamics) and $r = 4$ (chaotic dynamics) using $\gamma = 10\%$ and increase the noise standard deviation σ in steps of 0.01. Figure 3 shows that we can clearly distinguish between periodic dynamics and chaotic dynamics even for high noise levels. The bullets give the average determinism for a 100 time-series ensemble, while the error bars show the standard derivation of the ensemble. It should be noted that for these two extreme chases—periodic and chaotic—the bands are clearly separated for the whole σ range considered but to be on the safe side we would not recommend analyzing data with more than $\sigma = 0.2$. Nevertheless, our method is quite stable even if corrupted by measurement noise.

B. Rössler attractor

In order to mimic irregular sampling, we consider the continuous Rössler system:

$$\left(\frac{dx}{dt},\frac{dy}{dt},\frac{dz}{dt}\right) = [-y-z, x+ay, b+z(x-c)], \qquad (6)$$

where a, b, and c are parameters. In this paper we chose $a = 0.2$ and $c = 5.7$ and vary $b \in [0,1.4]$ with a resolution of $\Delta b = 0.01$. To achieve an irregular sampling, we use the maximum map $\tilde{Y}$ of the y component, which offers a natural way to get nonequidistant sampled event time series in the chaotic regime as well as in the windows of higher periodicity. For our investigation we generate a long time series via using the fourth-order Runge-Kutta method with $\Delta t = 0.01$ sampling rate. Then we neglect any transient behavior and consider 5000 maxima for each control parameter value b. Using a window size of 3500 time units we calculate TACTS with parameters $\lambda_{0,k}$ determined by Eqs. (2) for all b values and then optimize λ_S such that the cost distribution is normal.

In addition to the irregular sampling, real-world data have some measurement uncertainties. While we know that our method performs well even with measurement noise added to the dynamics (cf. Fig. 3), the Rössler system offers an additional opportunity to test for a different kind of uncertainty. As mentioned above, paleoclimate proxy records are often γ distributed in the time domain [4,5]. To let our data reflect this, we first create a cubic interpolated maximum time series resulting from $\tilde{Y}$ acting on y. Then we choose γ-distributed time events at which we sample the interpolated time series to create a new time series with higher uncertainty. The two-steps process is illustrated in Fig. 4 for two chosen skewness values. For our analysis we are using the skewness of the γ distribution as 0.3, 0.5, 1.0, and 2.0. Therefore, we generate four additional time series that we analyze by determining TACTS and determine our RQA measure DET with $\varepsilon = 0.05$.

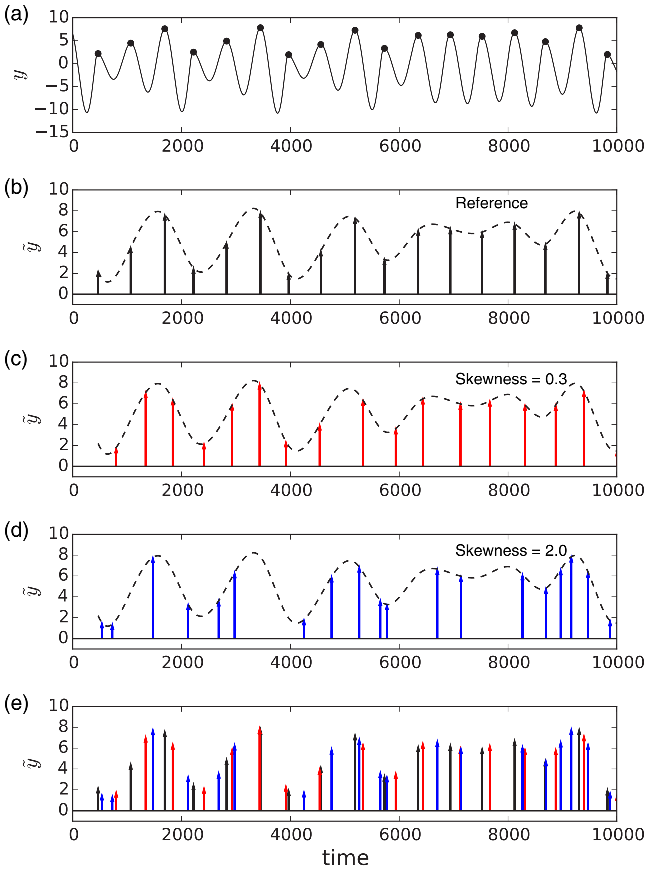

FIG. 4. (Color online) The time series of Rössler: Panel (a) shows the y component with maxima highlighted as bullets; panel (b) shows the result of the maximum map $\tilde{Y}$ acting on y; panel (c) shows the time series for a skewness of 0.3; and panel (d) shows the time series for a skewness of 2.0. In panels (b)–(d) we show the interpolated time series that we draw from as a black dashed line. Panel (e) offers a comparison between the different time series and highlights their irregularity.

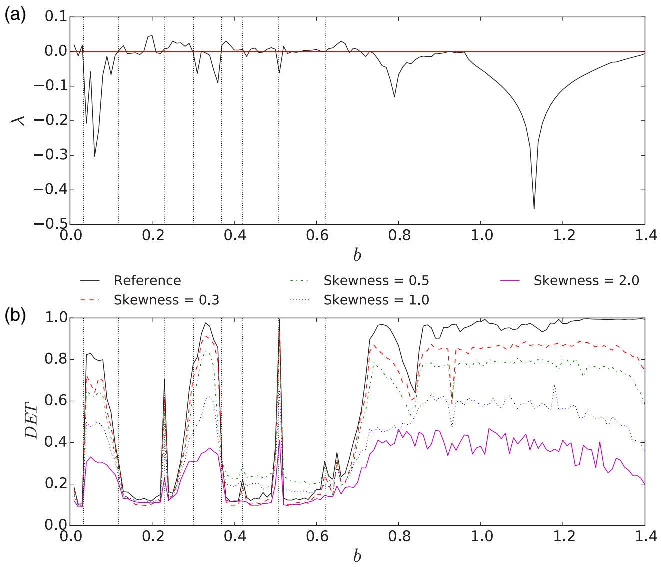

FIG. 5. (Color online) RQA analyses for Rössler map distance time series: In panel (a) the Lyapunov exponent λ is given as a reference over the whole b control-parameter range. In panel (b) the determinism DET determined from TACTS is shown for the maximum time series $\hat{Y}$ (black line) as a reference. The other shown data results from the γ-distributed time series with skewness 0.3 (red dashed line), skewness 0.5 (green dash-dot line), skewness 1.0 (blue dotted line), and skewness 2.0 (pink line).

The results for the five time series are shown in Fig. 5. Figure 5(a) shows the Lyapunov exponent calculated from the continuous sampled y component. In the panel below we see the data for the five different time series we considered. Clearly our technique is able to identify the dynamical regime changes (dotted lines in Fig. 5) for all data sets. Just like with increasing noise intensity in the logistic map the changes are not so pronounced anymore for higher skewness, but even for a skewness of 2.0 the chaotic regime can be clearly distinguished from the periodic dynamics and we are able to identify the regime changes associated with the periodic windows.

IV. APPLICATION TO PALEOCLIMATE RECORD

So far we have been testing the performance of our transformation-cost time-series method using prototypical models. While we have been trying to design these numerical examples as realistic as possible—including measurement noise and testing irregular γ-distributed sampling—in real applications the data might have further complications like multiplicative noise and time-dependent control parameters. Since we are particularly interested in paleoclimate applications, we choose a speleothem isotope δ^{18}O record from the Secret Cave at Gunung Mulu in Borneo, Indonesia [20]. This particular record is a proxy for the Indonesian-Australian monsoon, since δ^{18}O is an indicator for precipitation. We analyzed the last 62 000 years of this proxy record. Note that the full record is around 100 000 years long, but beyond 62 000 years the data is too sparse and contains too many gaps to give any useful information.

In the part of the record analyzed there are ~1200 data points. The time between two measurements follow a γ distribution and the skewness is 4.9. We use a segment size of ≈ 210 years to calculate our transformation-cost time series. The parameters $\lambda_{0,k}$ are determined by Eqs. (2) and we optimize $\lambda_S = 1.07$. To detect dynamical transitions in TACTS, we need to apply a slightly different form of the RQA [21], since in the paleoclimate data we expect a temporal variation of the control parameter. We therefore adopt a sliding window method and consider 30 data points of TACTS as our window size. Note that 30 data points in TACTS correspond to approximately 100 to 140 points in the original proxy record and cover about 6200 years in real time. The length of a window (~6200 years) is a suitable interval to detect the regime transitions in paleoclimatology. DET is determined for each window of the time series one by one as the window slides over the time series with 90% overlap. We used $\varepsilon = 20\%$ of the standard derivation of the data in the particular window. Not only does this method allow us to deal with data that shows regime changes due to control parameter variations, it also gives us a way to determine the statistical significance of DET via the method of bootstrapping as outlined in Ref. [21]. The basic idea is that the dynamics of the system does not change over time. The bootstrapping test will allow us to judge whether the found variability of the measure DET is significantly different from an unchanged dynamics, i.e., whether a regime transition occurs.

In Fig. 6 we present the results of our analysis together with the original proxy record. Outside the light red band DET can be considered to indicate a dynamics different from the

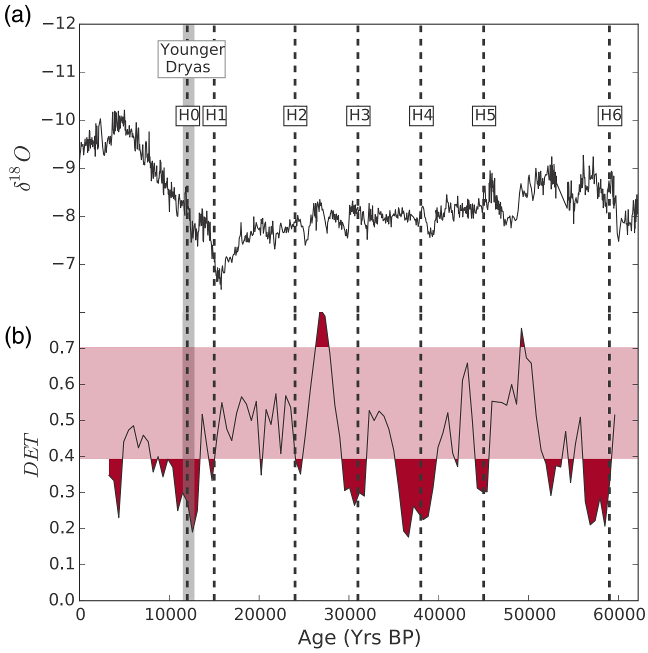

FIG. 6. (Color online) Upper graph: δ^{18}O record of the Secret Cave, Borneo; lower graph DET determined from the corresponding transformation-cost time series. Horizontal lines H1–H6 indicate the six Heinrich events, while the most recent line determines the Younger Dryas, a cold period in the northern hemisphere most likely caused by a collapse of the North American ice sheet. The light red band of the DET indicates the 90% confidence interval.

"normal behavior" with 90% confidence. As we can see, the RQA determinism indicates several distinct regime changes in the time series. Quite pronounced are the regime changes that coincides with the known Heinrich events (H1–H6). During these Heinrich events large quantities of fresh water were introduced into the Atlantic via melting ice-bergs [22], and it is apparent that these events impacted also on the Monsoon dynamics over the maritime continent [23], leading to very low DET. Similarly, the Younger Dryas, a period of cool climate in the northern hemisphere that might have been caused by the collapse of the North American ice sheet [24], is detected by our method. It is noteworthy that in the original work by Carolin *et al.* [20], H1–H6 was detected too, but the Younger Dryas coinciding with the Heinrich 0 event (H0) was not detected. Moreover, our method allows an objective, quantitative analysis, while Carolin *et al.* rely on the subjective method of matching extreme data occurrences with specific dates.

In our analysis we detect some other significant regime changes (see, for example, the high DET between H2 and H3 in Fig. 6) that have previously not been recognized. Similarly, we observe that while all Heinrich events were impacting on the climate significantly, the duration and strength with which they impacted on the monsoon over Indonesia varied according to our analysis. From a paleoclimatic point of view, the monsoon dynamics over the Maritime Continent is quite complex with cold surges from the north impacting on the local precipitation. In addition, changes in the landmasses due to rising sea levels generated the Borneo vortex, which is dominating in more recent times the monsoon [25]. We do not claim that our analysis of the speleothem record answers all questions and the aim of this paper is out of paleoclimatological scope. We are going to address these additional regime changes and the durations of the Heinrich events' impact in a more specific journal [26].

V. CONCLUSION

In this paper we have presented a novel method for analyzing irregularly sampled time series. This transformation-cost time-series (TACTS) method is based on determining similarities in time-series segments. The fundamental transformation of the segments follows several elementary steps, such as moving a data point in time or changing its amplitude. By analyzing the average sampling rate and the average amplitude one can determine the associated cost factors for these transformation steps. Moreover, as we have demonstrated the deletion and creation cost can be optimized relative to the first two costs. The advantage of our method is that the resulting transformation-cost time series is regularly sampled. Therefore, one is free to use a suitable time-series method for further analysis without the risk of data corruption arising from unsuitable interpolation methods.

Our extensive tests of the method have demonstrated that TACTS is useful even for extreme irregular sampling and in addition can deal with rather high measurement noise. It can be used in discrete and continuous systems and shows promising results when applied to real-world applications. In combination with recurrence plot analysis measures like the determinism DET our method can detect dynamical regime changes accurately. Especially in areas like paleoclimate, where often irregular sampling and parameter changes are common, our method provides a quantitative and objective way to analyze data and can guide scientists to previous hidden regime changes.

The systematic comparison of the effects of interpolation and TACTS on different time-series analysis results is the subject of an ongoing study that will be published in the future.

ACKNOWLEDGMENTS

We thank K.-H. Wyrwoll and F. McRobie from the School of Earth and Environment (UWA) for fruitful discussions. Moreover, we acknowledge for financial supports from TUBITAK under the 2214/A program (I.O.), from the Leibniz Association (WGL) under Grant No. SAW-2013-IZW-2542 (D.E.), as well as from the BMBF within the Potsdam Research Cluster for Georisk Analysis, Environmental Change and Sustainability (PROGRESS) Support Code No. 03IS2191B (N.M.).

[1] V. N. Livina and T. M. Lenton, Geophys. Res. Lett. **34**, L03712 (2007).

[2] L. L. Trulla, A. Giuliani, J. P. Zbilut, and C. L. Webber Jr., Phys. Lett. A **223**, 255 (1996).

[3] N. Malik, N. Marwan, Y. Zou, P. J. Mucha, and J. Kurths, Phys. Rev. E **89**, 062908 (2014).

[4] K. Rehfeld, N. Marwan, J. Heitzig, and J. Kurths, Nonlin. Proc. Geophys. **18**, 389 (2011).

[5] K. Rehfeld and J. Kurths, Climate Past **10**, 107 (2014).

[6] N. Marwan, M. C. Romano, M. Thiel, and J. Kurths, Phys. Rep. **438**, 237 (2007).

[7] J. D. Victor and K. P. Purpura, Network: Comput. Neural Syst. **8**, 127 (1997).

[8] Y. Hirata and K. Aihara, J. Neurosci. Methods **183**, 277 (2009).

[9] J. J. C. Michael T. Rosenstein and C. J. D. Luca, Physica D **65**, 117 (1993).

[10] X. Rodó and M.-A. Rodriguez-Arias, Climate Dynam. **27**, 441 (2006).

[11] H. Sakoe and S. Chiba, IEEE Trans. Acoust. Speech Signal Proc. **26**, 43 (1978).

[12] P. Grassberger, Phys. Lett. A **97**, 227 (1983).

[13] K. Rehfeld, N. Marwan, S. F. M. Breitenbach, and J. Kurths, Climate Dynam. **41**, 3 (2012).

[14] S. Suzuki, Y. Hirata, and K. Aihara, Int. J. Bifurcat. Chaos **20**, 3699 (2010).

[15] J.-P. Eckmann, S. Oliffson Kamphorst, and D. Ruelle, Europhys. Lett. **4**, 973 (1987).

[16] N. Marwan, J. F. Donges, Y. Zou, R. V. Donner, and J. Kurths, Phys. Lett. A **373**, 4246 (2009).

[17] D. Eroglu, N. Marwan, S. Prasad, and J. Kurths, Nonlin. Proc. Geophys. **21**, 1085 (2014).

[18] D. Eroglu, T. K. DM. Peron, N. Marwan, F. A. Rodrigues, L. F. Costa, M. Sebek, I. Z. Kiss, and J. Kurths, Phys. Rev. E **90**, 042919 (2014).

[19] N. Marwan, N. Wessel, U. Meyerfeldt, A. Schirdewan, and J. Kurths, Phys. Rev. E Statist. Nonlin. Soft Matter Phys. **66**, 1 (2002).

[20] S. A. Carolin, K. M. Cobb, J. F. Adkins, B. Clark, J. L. Conroy, S. Lejau, J. Malang, and A. A. Tuen, Science **340**, 1564 (2013).

[21] N. Marwan, S. Schinkel, and J. Kurths, Europhys. Lett. **101**, 20007 (2013).

[22] G. Bond, H. Heinrich, W. Broecker, L. Labeyrie, J. McManus, J. Andrews, S. Huon, R. Jantschik, S. Clasen, C. Simet, K. Tedesco, M. Klas, G. Bonani, and S. Ivy, Nature **360**, 245 (1992).

[23] F. S. R. Pausata, D. S. Battisti, K. H. Nisancioglu, and C. M. Bitz, Nature Geosci. **4**, 474 (2011).

[24] W. Berger, Nature **360**, 219 (1992).

[25] S. Koseki, T. Koh, and C. Teo, Atmospher. Chem. Phys. **14**, 4539 (2014).

[26] D. Eroglu, T. Stemler, F. McRobie, K.-H. Wyrwoll, N. Marwan, and J. Kurths (unpublished).