




RESEARCH ARTICLE | OCTOBER 02 2023

Interpolation and sampling effects on recurrence quantification measures

Special Collection: [Nonlinear dynamics, synchronization and networks: Dedicated to Jürgen Kurths' 70th birthday](#)

Nils Antary ; Martin H. Trauth ; Norbert Marwan 



Chaos 33, 103105 (2023)

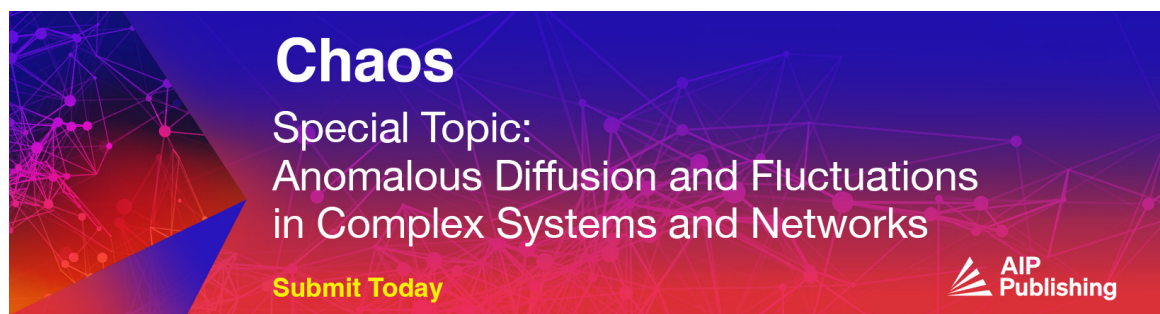
<https://doi.org/10.1063/5.0167413>



View
Online




Export
Citation



Chaos

Special Topic:
Anomalous Diffusion and Fluctuations
in Complex Systems and Networks

[Submit Today](#)



Interpolation and sampling effects on recurrence quantification measures

Cite as: Chaos 33, 103105 (2023); doi: 10.1063/5.0167413

Submitted: 12 July 2023 · Accepted: 6 September 2023 ·

Published Online: 2 October 2023



View Online



Export Citation



CrossMark

Nils Antary,^{1,2,a)}  Martin H. Trauth,³  and Norbert Marwan^{1,3} 

AFFILIATIONS

¹Potsdam Institute for Climate Impact Research (PIK), Member of the Leibniz Association, 14473 Potsdam, Germany

²Institute for Theoretical Physics, University of Leipzig, 04081 Leipzig, Germany

³Institute of Geosciences, University of Potsdam, Karl-Liebknecht-Straße 24–25, 14476 Potsdam, Germany

Note: This paper is part of the Focus Issue on Nonlinear dynamics, synchronization and networks: Dedicated to Juergen Kurths' 70th birthday.

^{a)}Author to whom correspondence should be addressed: nantary@protonmail.com

ABSTRACT

The recurrence plot and the recurrence quantification analysis (RQA) are well-established methods for the analysis of data from complex systems. They provide important insights into the nature of the dynamics, periodicity, regime changes, and many more. These methods are used in different fields of research, such as finance, engineering, life, and earth science. To use them, the data have usually to be uniformly sampled, posing difficulties in investigations that provide non-uniformly sampled data, as typical in medical data (e.g., heart-beat based measurements), paleoclimate archives (such as sediment cores or stalagmites), or astrophysics (supernova or pulsar observations). One frequently used solution is interpolation to generate uniform time series. However, this preprocessing step can introduce bias to the RQA measures, particularly those that rely on the diagonal or vertical line structure in the recurrence plot. Using prototypical model systems, we systematically analyze differences in the RQA measure average diagonal line length for data with different sampling and interpolation. For real data, we show that the course of this measure strongly depends on the choice of the sampling rate for interpolation. Furthermore, we suggest a correction scheme, which is capable of correcting the bias introduced by the preprocessing step if the interpolation ratio is an integer.

© 2023 Author(s). All article content, except where otherwise noted, is licensed under a Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>). <https://doi.org/10.1063/5.0167413>

Almost all natural systems are non-linear systems, often with multiple dimensions. The analysis of the possibly rich dynamics of such systems requires advanced methods. One of these is the recurrence plot and the associated recurrence quantification analysis, which are, among other things, used to investigate the nature of the dynamics, periodicity, or to detect regime transitions. As with most other methods, this method was developed for uniformly sampled data. This, however, restricts the use of the method. Some data from, for example, astrophysics or paleoclimate cannot be analyzed straightforwardly. To circumvent this restriction, one used approach is to interpolate the data to a constant sampling rate. We show that the differences in the sampling rate together with the subsequent interpolation can lead to strong deviations, and it is not recommendable to take this approach without further consideration. We, thus, propose

a correction scheme that can, with some limitations, correct these deviations.

I. INTRODUCTION

The analysis of data from complex real-world systems creates the basis for many different fields in science, such as finance, engineering, life, and earth science. For many kinds of data, standard measures, such as mean, standard deviation, or higher moments, are not sufficient to capture the details of their dynamics. For this purpose, methods from complex system science are more appropriate, such as Lyapunov exponents,¹ complex networks,² symbolic dynamics,³ or recurrence analysis.⁴ Recurrence analysis provides a set of tools, e.g., for studying synchronization, classifying different

types of dynamics, or detecting regime transitions.⁵ The increasing popularity of this framework is reflected by the lively methodical development and the growing number of applications in many scientific disciplines.⁶

In several fields, data are only available on a non-equidistant time axis; for instance, measurements based on heartbeats are timed by the rhythm of the heart,^{7,8} which gives a natural varying timescale, and this leads to the necessity of interpolation when compared with other variables that form the cardiorespiratory system, such as the blood pressure.⁹ Time series of pulsars and rotating stars are obtained through non-equidistant observations,^{10,11} and paleoclimate data are hampered by the non-constant sedimentation rate, resulting in non-equidistant sampling.¹² Such unevenly sampled data are a challenge for most time series analysis techniques, which usually require equidistant sampling points. One solution is to transform the time series to an equidistantly sampled one using interpolation or other techniques (e.g., transformation cost¹³). Other approaches try to modify the time series analysis tools to be directly applicable to uneven time series, such as Lomb–Scargle periodogram,¹⁴ kernel-based correlation,^{15,16} or edit distance-based recurrence analysis.¹⁷ However, interpolation is still a widely used technique, although it can cause serious bias in the results (overemphasizing the lower frequencies). Although this interpolation effect is known for several time series analysis techniques,¹⁸ it is not yet systematically considered for recurrence analysis and, thus, can lead to wrong interpretations and conclusions.

Our focus in this study is the effect of non-uniform sampling in paleoclimate time series on the results of recurrence analysis, as it has recently gained attention for its potential to address various research questions in geosciences.^{19–24} A prominent example is the investigation of transitions in the paleoclimate.^{25–29} Changes in the recurrence properties, mainly based on changes in the distribution of the diagonal line structures in recurrence plots (related to determinism or predictability of the underlying dynamics), can be used to identify regime changes. Paleoclimate data are usually retrieved from specific geological archives, e.g., marine and lake sediments, tree rings, ice cores, or stalagmites. The climate information is stored during the growth (deposition) of these archives. Since the growth or sedimentation rate can differ over time, but the sampling procedure of such archives usually uses an equidistant sampling scheme, and the final time series are usually unevenly sampled in time.¹² Using interpolation without due consideration before conducting recurrence analysis can lead to bias in the distribution of diagonal line structures in the recurrence plots, finally resulting in erroneous conclusions.

In this work, we show the effects of different and changing sampling times and subsequent interpolation on the recurrence quantification analysis using different types of model systems as well as a real paleoclimate example.

We further suggest an approach to estimate quantitatively these effects and provide a correction scheme. Although we focus here on applications in paleoclimate research, the correction scheme can be analogously applied on other research questions, e.g., in astrophysics or physiology.

This work is structured as follows: In Sec. II, the recurrence plot and recurrence plot measures are introduced. In Sec. III, the used models and methods are given. Then, we present the measured

effects for the simulated systems. Afterward, in Sec. IV, we present the derivation as well as the evaluation of our correction scheme. In Sec. V, a real-world example is considered. We conclude our findings in Sec. VI.

II. PRINCIPLES OF RECURRENCE PLOTS (RPs) AND RECURRENCE QUANTIFICATION ANALYSIS (RQA)

The recurrence plot (RP) is based on the fundamental principle that a dynamical system will always return to a state arbitrarily close to its initial or any other state.⁴ This fact is used to simplify the generally multidimensional phase space trajectory to a matrix containing only zeros and ones. This method can be applied to any series of states in a given phase space $\{\vec{x}_i | i = 1, \dots, N\}$, where \vec{x}_i is the phase space vector at the time step i (corresponding to time $t = i \cdot dt$ and dt the sampling time) and N the number of considered states (or the length of the time series). If we do not have access to the full phase space vector, it can be reconstructed using time delay embedding.^{30,31} To create a matrix representation of the recurrences of the data, the phase space distance from all pairs of data points is calculated using a suitable norm, represented by the distance matrix

$$D_{ij} = \|\vec{x}_i - \vec{x}_j\|, \quad i, j = 1, \dots, N. \quad (1)$$

If not stated otherwise, we use the Euler norm. A recurrence is finally defined as having the pairwise distance D_{ij} between the states smaller than a specified recurrence threshold ε ; i.e., the recurrence matrix is derived from the distance matrix \mathbf{D} by applying the threshold ε , leading to the binary recurrence matrix \mathbf{R} with its elements

$$R_{ij}(\varepsilon) = \Theta(\varepsilon - D_{ij}), \quad (2)$$

with Θ being the Heaviside function and ε being the recurrence threshold.⁴

The recurrence matrix \mathbf{R} can be displayed as a plot, where all the ones ($R_{ij} = 1$) are marked by points, and such an entry is, therefore, called a “recurrence point” or a “1-point.” In contrast, the points $R_{ij} = 0$ are called a “non-recurrence point” or a “0-point.” This plot is then called a recurrence plot (RP) and can be visually interpreted because it expresses rich patterns characteristic of specific dynamics.^{4,32}

Although the first visual impression gives some important hints, further (quantitative) analysis is often useful. For this purpose, the recurrence quantification analysis (RQA) was introduced.^{7,33,34} The recurrence rate RR , the fraction of recurrence points in the RP, gives a quantification of how often the system returns to the same region in the phase space. Most of the other measures in RQA rely mainly on the length distribution $P_l(l)$ of either diagonal or vertical lines (formed by the recurrence points) visible in the RP. Such a diagonal line is denoted here as a “recurrence line” or a “1-line.” The idea behind the study of diagonal recurrence lines is that their length corresponds to the time the system evolves similarly compared to the other times the system visited the same region in the phase space. Therefore, the fraction of 1-points that form diagonal 1-lines in the RP is a heuristic measure of how deterministic the system is,

$$DET = \frac{\sum_{l=l_{\min}}^N l P_l(l)}{\sum_{l=1}^N l P_l(l)}, \quad (3)$$

with $P_l(l)$ being the probability to find a 1-line with exact length l , l_{\min} a chosen minimal line length, and N the maximal line length (equal to data length). This measure is called *determinism* and ranges between 0 and 1. It becomes 1 if all recurrence points belong to diagonal lines equal or longer than l_{\min} and 0 if no recurrence point does. This measure is widely used to identify regime transitions, e.g., in paleoclimate studies.^{27–29,35,36}

The calculation of *DET* is based on the distribution of diagonal line lengths. For the sake of simplicity, here, we focus on the average diagonal length,

$$L = \sum_{l=1}^N l P_l(l). \quad (4)$$

In contrast to the standard definition of this measure,⁴ we consider here all line lengths, including such of length 1. This simplification allows us to construct correction schema for possible sampling and interpolation effects. This measure is related to the prediction time. However, it depends on the temporal resolution of the system. For Gaussian white noise, L is $1/(1 - RR)$ and, therefore, for small RR close to one, whereas for perfectly periodic systems, it should theoretically be infinite but is limited by the RP size and boundary.³⁷ L is very sensitive to noise because noise causes random interruptions in the diagonal lines. In paleoclimate studies, measures quantifying the diagonal lines are often interpreted in terms of the predictability of the climate.^{29,35,36}

In many cases, the temporal variation of the RQA measures is of interest as they can reveal changes or transitions in the system's dynamics, such as when the system is approaching a tipping point. To determine such changes, the sliding window approach is used. The time series is partitioned into smaller segments (windows) of a predetermined length, which may overlap with one another. Then, the RP and the RQA measures are calculated for every window separately. The distance between the start points of two consecutive windows is called the window step size, and the size of a window is the window size. It is important to determine which time point is assigned to each window. In this study, we just take the starting point of the window so that all points used to calculate the measure are in the interval after this time point.

III. INTERPOLATION EFFECT ON RECURRENCE ANALYSIS OF MODEL SYSTEMS

To demonstrate the interpolation effect, we use two model systems to generate prototypical data. We use an autoregressive model, where the course of the data is stochastically driven, and a Rössler³⁸ system, which is a typical non-linear system, fully described by three non-linear ordinary differential equations.

To quantify the deviation in the L measure due to the difference in the sampling rate and subsequent interpolation, we compare L^{ref} calculated from a reference series without interpolation with L^{int} calculated from the interpolated series with the same temporal resolution as the reference series by taking the ratio $L^{\text{int}}/L^{\text{ref}}$. Ratios greater than 1 indicate an increase of the measure due to the interpolation and ratios smaller than 1 a decrease. This could then either lead to an over- or underestimation of an L^{ref} value when analyzing the interpolated series.

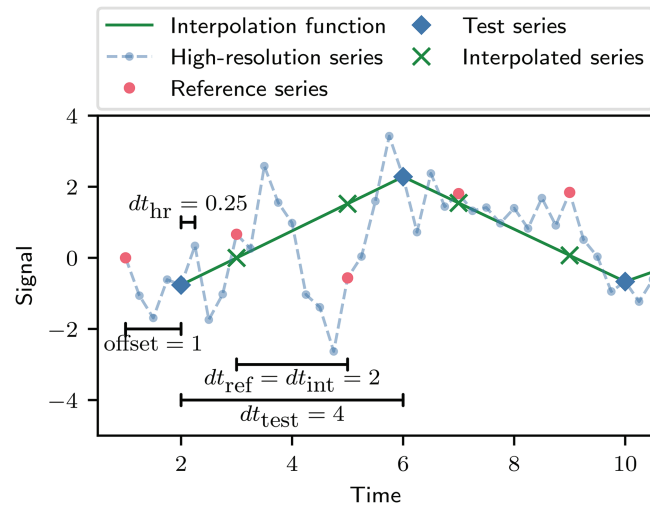


FIG. 1. Schema illustrating the different series as well as the offset and interpolation.

The effect depends on the considered system, the interpolation ratio, and the offset. The interpolation ratio r is the ratio of the sampling time of the time series before and after interpolation. The offset is the time difference between the first value in the time series before and after interpolation (Fig. 1).

An interpolation ratio greater than 1 increases the total number of points. For integer interpolation ratios, every r th point in the interpolated series lines up with the underlying series only if the offset is zero.

A. Systems

We use the two different model systems for our study and generate the data using the following equations. To remove transient behavior from the initial conditions to some kind of stable dynamics, the first part of every time series is discarded.

1. Autoregressive model

The time series is generated iteratively according to the rule

$$x_{i+1} = ax_i + \sigma \xi_i, \quad (5)$$

where ξ is Gaussian white noise (mean zero and standard deviation one) and a and σ are free parameters, defining the auto-correlation (memory) and the impact of the noise. We used a parameter setting of $a = 0.99$ and $\sigma = 1$. Three series are generated for the analysis with each 2501 data points. The initial conditions are drawn from a normal distribution, and the first 100 points are discarded. The noise and the initial conditions are generated with a random number generator with three different seeds to make the results reproducible.

2. Rössler model

The Rössler system is fully described by the differential equations³⁸

$$\begin{aligned} \dot{x} &= -y - z, \\ \dot{y} &= x + ay, \\ \dot{z} &= b + z(x - c). \end{aligned} \tag{6}$$

We use the standard parameters $a = 0.15$, $b = 0.2$, $c = 10$ and a temporal resolution of $dt = 0.01$. To approximately solve these equations, we used the Euler method and the initial condition $(x = 15, y = 0, z = 0)$. The obtained series has 12 505 data points, and the first 500 points are discarded. Afterward, only every fifth data point is used so that the series of both systems have 2401 points.

B. Method

To mimic the effect of different sampling times and subsequent interpolation, we consider high-resolution series, representing the “true” dynamics. To obtain series with different temporal resolution, multiple downsampled versions are constructed. One downsampled time series is chosen as reference series and analyzed without interpolation so that the results are unchanged and describe the dynamics of the system. All other downsampled series are called the test series (see Fig. 2). In the analysis, the offset for the reference series is chosen to be 0 because only the difference in the offset between the reference and the test series is important. This gives the following series:

$$\begin{aligned} \text{High-resolution series: } & \{\bar{x}_i | i = 0, 1, 2, \dots, N_{hr} - 1\}, \\ \text{Reference series: } & \{\bar{x}_i | i = 0, 1 \cdot d_{ref}, 2 \cdot d_{ref}, \dots, (N_{ref} - 1) \cdot d_{ref}\}, \\ \text{Test series: } & \{\bar{x}_i | i = k_{test,n}, 1 \cdot d_{test,n} + k_{test,n}, 2 \cdot d_{test,n} + k_{test,n}, \dots, (N_{test,n} - 1) \cdot d_{test,n} + k_{test,n}\}, \end{aligned}$$

where N_{hr} , N_{ref} , and $N_{test,n}$ are the lengths of the different series, d_{ref} is the downsampling factor of the reference series, and $d_{test,n}$ and $k_{test,n}$ are the downsampling factors and offsets of the different test series, where n numbers the different test series. The corresponding times are given by $t_i = i \cdot dt$. We generate interpolation functions for every test series using linear, quadratic spline, cubic spline, and pchip interpolation. To obtain the interpolated series, we evaluated these interpolation functions at the same time points as the reference series (see Fig. 1).

Using the reference instead of the high-resolution time series as a comparison, it is possible to study non-integer interpolation ratios, offsets, and also interpolation ratios smaller than one.

The interpolation ratios are given as

$$r_n = \frac{dt_{test,n}}{dt_{int}} = \frac{dt_{test,n}}{dt_{ref}} = \frac{d_{test,n} \cdot dt_{hr}}{d_{ref} \cdot dt_{hr}} = \frac{d_{test,n}}{d_{ref}}, \tag{7}$$

where dt_{hr} , dt_{ref} , and $dt_{test,n}$ are the different sampling times from the series defined above and dt_{int} is the sampling time of the interpolated series.

The RP and the RQA measure L are calculated for the reference and for every interpolated series using the same recurrence threshold ε . L^{ref} is the measure obtained for the reference series and L^{int} for the interpolated series. The differences in the RQA measure are due to the different sampling times and the interpolation. We can compare the results from all interpolated series to the reference series by computing all ratios L^{int}/L^{ref} .

Using this procedure, we mimic the typical sampling bias, e.g., in paleoclimate studies, where sliding windows with different sampling are all interpolated to the same reference sampling and the hidden “true” dynamics is the continuous nature and is not accessible.

C. Results

To illustrate the deviations in the RP between an interpolated series and a reference series, we consider a short time series of an arbitrary autoregressive process and downsample it by a factor of four before interpolating it linearly back to the original size. This procedure leads to an interpolation ratio of $r = 4$ and no offset. Here, the high-resolution time series corresponds to the reference series. Comparing the RP of the interpolated time series with the RP of the reference time series, we find some differences (Fig. 3). Points that are 1-points in both RPs are black, points that are only 1-points in the reference RP are red, and points that are only 1-points in the interpolated RP blue. The coarse structure is preserved under the

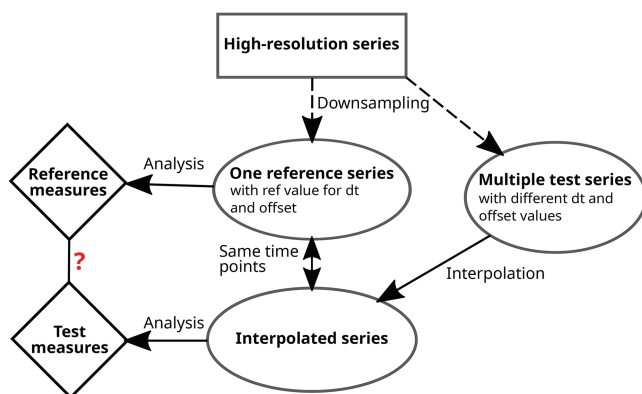


FIG. 2. Scheme illustrating the method used to investigate the change in the RQA measures due to the difference in the sampling time, offset, and interpolation.

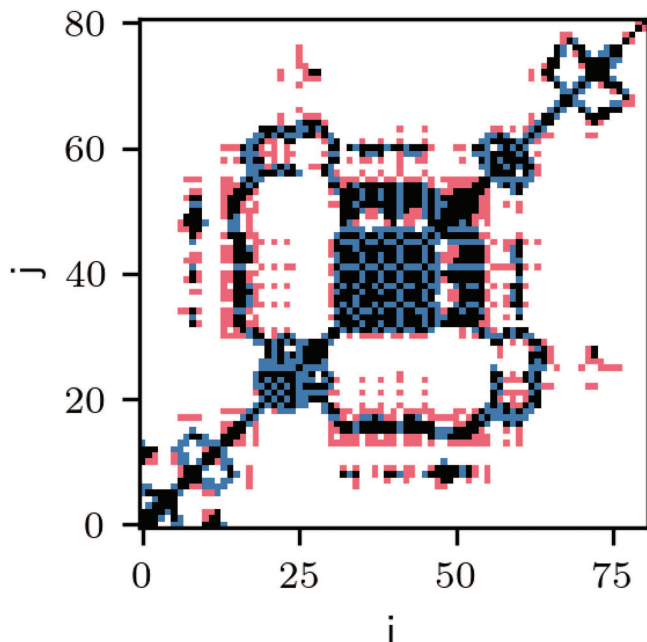


FIG. 3. Difference of interpolated and reference recurrence plots of an arbitrary AR process. Black pixel mark 1-points in both RPs; red pixel mark 1-points in reference RP, which 0-points in interpolated RP; and blue pixel mark 0-points in reference RP, which are 1-points in interpolated RP. The interpolation ratio is $r = 4$ and offset 0.

downsampling and interpolation; however, on a smaller scale, many short 1-lines are missing in the interpolated RP, while other 1-lines are merged into greater recurrence structures. Both effects lead to an increased average 1-line length.

To quantitatively study the effects on the average 1-line length, we use the method described in Sec. III B. For every model system, the high-resolution series is created according to Sec. III A. From this one, reference and multiple test series are created by downsampling with different offsets and downsampling factors $d_{\text{test},n}$. We use downsampling factors between 1 and 50 and for every downsampling factor five different offsets, if possible. For series with a downsampling factor smaller than five, the number of different offsets is limited by the downsampling factor.

The reference series are obtained with a downsampling factor $d_{\text{ref}} = 5$ and no offset. The high-resolution series have a length of 2401. The reference series have, therefore, 481 data points. The recurrence thresholds for the systems are chosen so that the RP of the reference series has a recurrence rate of 10%. The recurrence thresholds as well as the L^{ref} measure are given for the Roessler and the three autoregressive systems in Table I. The number in the name of the autoregressive systems (AR-42, AR-43, and AR-44) states the used seed for the random module of the Python library NumPy. With the three different autoregressive systems, we can check whether the results are changed for other realizations of the noise. For every system, the ratio $L^{\text{int}}/L^{\text{ref}}$ is calculated for every interpolation ratio, interpolation method, and offset separately. To

TABLE I. Properties of the four different reference series.

System name	Recurrence threshold ε	Reference L^{ref}
Roessler	5.06	24.6
AR-42	1.13	1.44
AR-43	1.05	1.36
AR-44	1.26	1.46

simplify the data, the mean and the standard deviation are calculated over the different offsets, and for the autoregressive systems, also over the three different realizations. The result is, therefore, only dependent on the kind of the system, the interpolation ratio, and the interpolation method.

1. Roessler system

For the Roessler system, L^{ref} and L^{int} values are almost equal for interpolation ratios smaller than 1.5 (Fig. 4 left). For a larger interpolation ratio, the L^{int} values for the different interpolated series are smaller than L^{ref} and decrease with an increasing interpolation ratio. The quadratic and the cubic spline interpolation lead to very similar results, whereas the linear and the pchip interpolation create the strongest deviation. The differences caused by the offset values are negligible for the linear interpolation, except at the integer ratios 2, 3, and 4. An explanation for this behavior is given in Sec. IV. Additionally, the same analysis was done with the maximum norm instead of the Euler norm. The results are qualitatively similar (see Appendix B).

2. Autoregressive process

For the interpolated series from the autoregressive systems, L^{int} increases with growing interpolation ratios for all methods (Fig. 4, right). L^{int} calculated from the linearly interpolated series shows the strongest deviation from the reference series. The effect of different realizations and offsets is similar for all considered interpolation methods.

To investigate the deviation caused by different offsets, $L^{\text{int}}/L^{\text{ref}}$ is calculated without offset and compared with the mean over all offsets. We find that the deviation is only different for integer interpolation ratios (Fig. 5). This is as expected because only for integer interpolation ratios, an offset determines whether all time points of the test series are aligned (without offset) or misaligned (with offset) to the interpolation time points (see Fig. 1). The $L^{\text{int}}/L^{\text{ref}}$ value is smaller without an offset, which means that the deviation of L^{int} from L^{ref} is smaller.

This result shows that interpolation can have a remarkable impact even if the interpolation ratio is 1 if there is some time offset between the points of the interpolated and test series. This situation commonly arises when the sampling time varies, causing misalignment between the interpolated series and the data. Even in portions of the data where the sampling time in the interpolated series is equal to the data's sampling time, the exact timings of the different series can be misaligned.

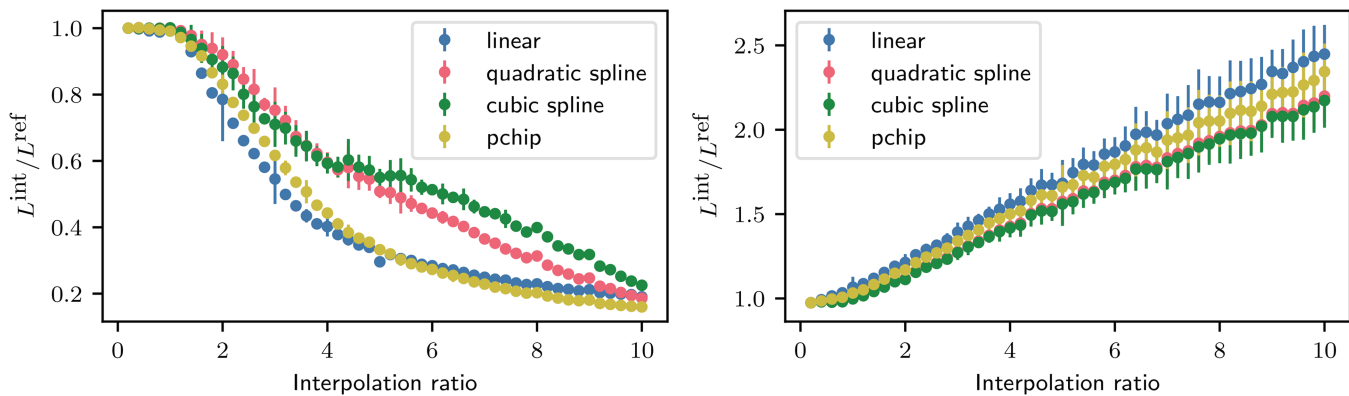


FIG. 4. Relative deviation of L between interpolated and reference series for the Rössler system (left) and three autoregressive systems (AR) (right). The data points represent the mean over the different systems (only for AR) and offsets, and the error bars give the standard deviation.

IV. CORRECTION SCHEME

Knowing the specific effect of interpolation on RQA measure L , it becomes apparent that a correction scheme aimed at mitigating this interpolation effect would be both desirable and feasible.

A. Method

To construct a correction scheme, which models the deviation in the L measure between an interpolated series to the reference series, it is necessary to account for the influences of interpolation and the differences in the sampling rate.

L can be calculated using the total number of diagonal 1-lines (i.e., lines consisting of values 1),

$$L = \frac{N_r}{N_l} = \frac{N^2 \cdot RR}{N_l}, \quad (8)$$

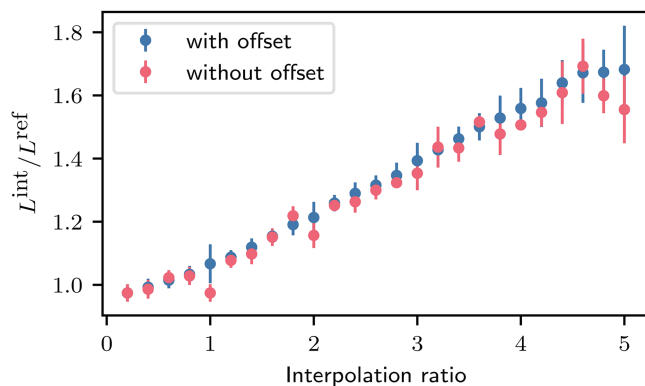


FIG. 5. Relative deviation of L between interpolated and reference series for three autoregressive systems (same as Fig. 4, right). The data points represent the mean over the different systems with (blue) and without (red) different offsets, and the error bars give the standard deviation.

where N_r is the total number of 1-points in the RP, N_l is the number of 1-lines (including the single points, i.e., lines with $l = 1$), N is the length of the series, and, therefore, N^2 is the total number of points in the RP, and RR is the recurrence rate.

The total number of points N^2 in the RP generated from the reference series (\mathbf{R}^{ref}) and the interpolated one (\mathbf{R}^{int}) is equal. In all investigated examples, the recurrence rate RR is also very similar; therefore, the deviation in L can be derived from the difference in the number of 1-lines N_l . As we will show, the number can differ, because

1. separated 1-lines in \mathbf{R}^{ref} can be connected in \mathbf{R}^{int} ,
2. 1-lines in \mathbf{R}^{ref} can be missing in \mathbf{R}^{int} , and
3. parts of 0-lines in \mathbf{R}^{ref} can form 1-lines in \mathbf{R}^{int} .

(1) and (2) lead to fewer 1-lines in \mathbf{R}^{int} and a greater L compared to \mathbf{R}^{ref} . (3) leads to more 1-lines and a smaller L .

To tackle the problem, we restrict ourselves to the following case:

- There is an integer interpolation ratio between the test series and the interpolation series and no offset.

This means that the test series consists of every r th data point from the reference series and is, therefore, a downsampled version of it. The downsampling results in a RP (\mathbf{R}^{test}) consisting of every r th point in every r th row of the reference RP (\mathbf{R}^{ref}) (with r being the interpolation ratio). The same is true for the diagonals: The i th diagonal in \mathbf{R}^{test} consists of every r th point from the $(r \cdot i)$ th diagonal of \mathbf{R}^{ref} . We call these points anchor points and these diagonals anchor diagonals (Fig. 6). To calculate the relative change in the number of 1-lines, we restrict ourselves to these diagonals.

At first, we calculate the difference in the number of 1-lines, due to the different sampling, by comparing the structures in \mathbf{R}^{ref} and \mathbf{R}^{test} . There are more 1-lines in \mathbf{R}^{ref} if there are more 1-lines starting between two anchor points than depicted by them. We refer to the sequences of 0- and 1-points between two anchor points as intervals. Only if the first anchor point is a 0-point and the second one is a 1-point, there is a starting point in \mathbf{R}^{test} ; all further starting points

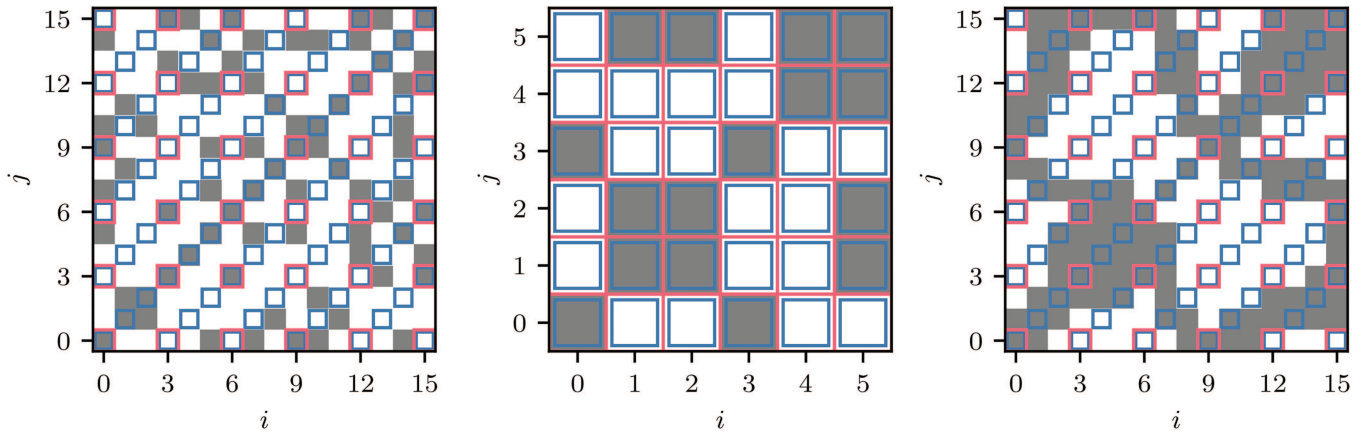


FIG. 6. Recurrence plots of some arbitrary reference signal (left), the test signal (middle), and the interpolated signal (right). The red boxes indicate the anchor points, which are identical in all plots. Blue boxes indicate all points belonging to the anchor diagonals. The interpolation ratio r is 3.

in \mathbf{R}^{ref} are additional. To quantify this effect, we can calculate the mean number of 1-line starting points. This is calculated separately for intervals following a 1- or a 0-anchor point,

$$\langle \Delta N_i \rangle_1 = \sum_{\Delta N_i=0}^r \Delta N_i \cdot P_{\Delta N_i,1}(\Delta N_i), \quad (9)$$

$$\langle \Delta N_i \rangle_0 = \sum_{\Delta N_i=0}^r \Delta N_i \cdot P_{\Delta N_i,0}(\Delta N_i), \quad (10)$$

where $\langle \Delta N_i \rangle_1$ is the mean number of additional 1-line starting points per interval, for intervals following 1-points, and $\langle \Delta N_i \rangle_0$ for intervals following 0-points. $P_{\Delta N_i,1}(\Delta N_i)$ and $P_{\Delta N_i,0}(\Delta N_i)$ are the probabilities to find intervals with ΔN_i 1-lines starting points, starting on either 1-points or 0-points.

To find these probabilities, we calculate the length distribution for sequences of alternating 0- and 1-lines, which have ΔN_i 1-lines starting points and calculate the probability that they fit between two anchor points. Such a sequence between two 1-points has a length between

$$l_{\Delta N_i,1,1} = l_s + \sum_{i=1}^{\Delta N} d_i + \sum_{i=1}^{\Delta N-1} l_i \quad \text{and} \quad (11)$$

$$l_{\Delta N_i,1,1}^* = l_s + \sum_{i=1}^{\Delta N} (d_i + l_i). \quad (12)$$

l_s is the start length, which is the number of 1-points after a random point on a random 1-line and d_i and l_i are the length of random 0-lines and 1-lines. Here, the second equation includes the 1-line containing the second 1-point and the first does not [example for $\Delta N_i = 2$ in Fig. 7(a)]. A sequence between a 1-point and a 0-point, which has ΔN_i 1-line starting points, has a length between [example

for $\Delta N_i = 1$ in Fig. 7(b)]

$$l_{\Delta N_i,1,0} = l_s + \sum_{i=1}^{\Delta N} (d_i + l_i) \quad (13)$$

and

$$l_{\Delta N_i,1,0}^* = l_s + \sum_{i=1}^{\Delta N+1} d_i + \sum_{i=1}^{\Delta N} l_i. \quad (14)$$

To get the probability that such sequence fits between a 1-anchor point and a random second anchor point, we add up probabilities for a 1- and a 0-point as a second anchor. In both cases, the probability is derived from the probability that the sequence is shorter than the interval length, multiplied by the probability of reaching the next anchor point when considering the next line together. The second factor is always a conditional probability because both lengths are not independent of each other. The distance between two anchor points and, therefore, the interval length is the same as the interpolation ratio r . The underlying probability distributions are given later. In total, we get

$$P_{\Delta N_i,1}(\Delta N_i) = P(l_{\Delta N_i,1,1} < r) \cdot P(l_{\Delta N_i,1,1}^* \geq r | l_{\Delta N_i,1,1} < r) + P(l_{\Delta N_i,1,0} < r) \cdot P(l_{\Delta N_i,1,0}^* \geq r | l_{\Delta N_i,1,0} < r), \quad (15)$$

where r is the interpolation ratio and $P(X < r)$ are cumulative probabilities, which can be calculated from the probability distributions: $P(X < r) = \sum_{X=0}^{r-1} P_X(X)$, where $P_X(X)$ are the probability distributions of the different sequence lengths.

The first two rows give the probability that if the first anchor point is a 1-point that there are ΔN 0-lines and $(\Delta N - 1)$ 1-lines before the next anchor point, which is also a 1-point. The third and the fourth row calculate the probability that there are ΔN 0-lines and ΔN 1-lines before the next anchor point, which is a 0-point. In both cases, the total interval has ΔN more 1-lines than the interval

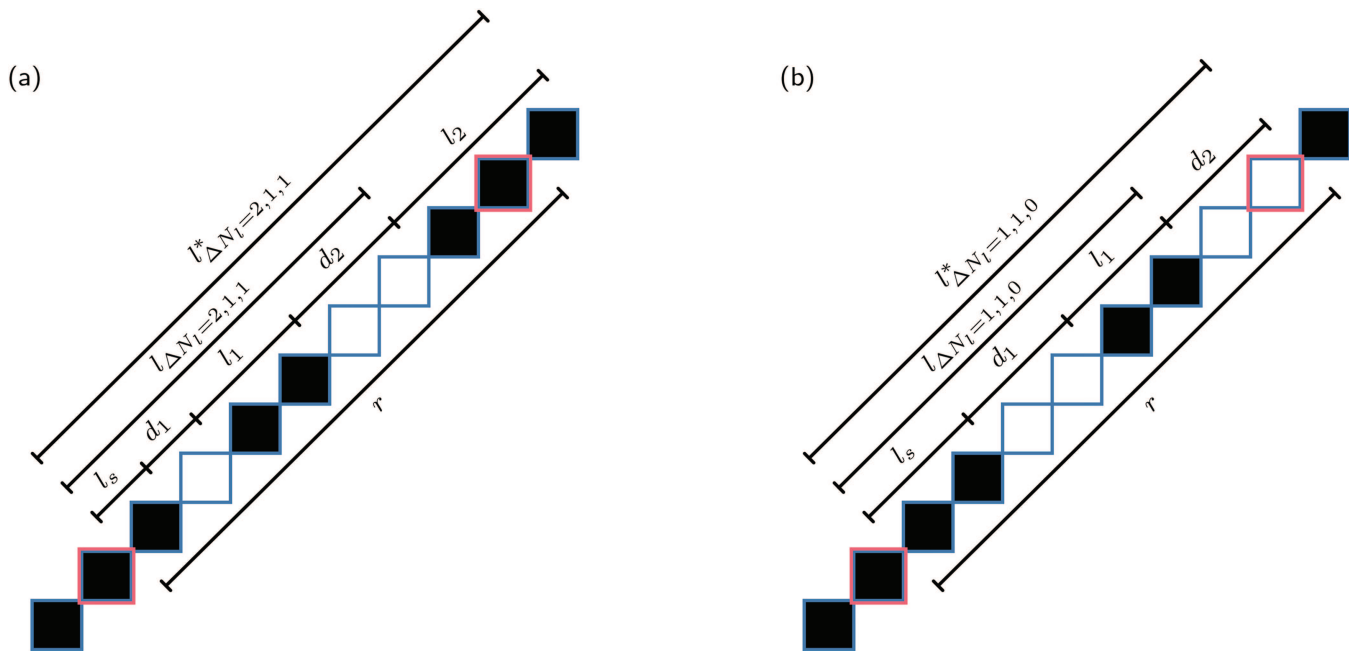


FIG. 7. Example intervals (a) between two 1-anchor points, showing one possible configuration with two additional 1-line starting points. (b) Between one 1-anchor point and one 0-anchor point, showing one possible configuration with one additional 1-line starting point.

downsampled to the two anchor points. This is explicitly given for all possible configurations for one example interval in Appendix C.

The first and third row are calculated from the probability distribution of the sum of the given random variables [Eqs. (11) and (13)]. For our calculation, we assume that the probabilities for the length of the consecutive 1- and 0-lines are independent. This gives

$$P_{l_{\Delta N_f,1,1}}(l_{\Delta N_f,1,1}) = (P_{l_s} * [P_d * P_l]_{x(\Delta N_f-1)} * P_d)(l_{\Delta N_f,1,1}), \quad (16)$$

$$P_{l_{\Delta N_f,1,0}}(l_{\Delta N_f,1,0}) = (P_{l_s} * [P_d * P_l]_{x(\Delta N_f)})(l_{\Delta N_f,1,0}), \quad (17)$$

where $*$ represents a convolution and $[P_d * P_l]_{x(M)}$ indicates that this part of the equation is repeated M times. $P_{l_s}(l_s)$, $P_d(d)$, and $P_l(l)$ are

the probability distributions for the start length, the 0-lines, and the 1-lines. $P_d(d)$ and $P_l(l)$ have to be known and are only accessible from the 0- and 1-line length histograms of the reference recurrence plot. How to estimate these when the reference series is not known is discussed later. The probability distribution for the start length l_s is (see Appendix A)

$$P_{l_s}(l_s) = \frac{\sum_{l=l_s+1}^{\infty} P_l(l)}{\sum_{l=1}^{\infty} l \cdot P_l(l)}; \quad (18)$$

for the conditional probabilities in row two, we first need to modify the probability from before so that the condition is fulfilled,

$$P_{l_{\Delta N_f,1,1}}^*(l_{\Delta N_f,1,1}) = \frac{1}{\sum_{l_{\Delta N_f,1,1}=0}^{r-1} P_{l_{\Delta N_f,1,1}}(l_{\Delta N_f,1,1})} \times \begin{cases} P_{l_{\Delta N_f,1,1}}(l_{\Delta N_f,1,1}) & l_{\Delta N_f,1,1} < r \\ 0 & \text{else,} \end{cases} \quad (19)$$

where the part in front of the curly bracket is the normalization factor. This can now be used to calculate

$$P(l_{\Delta N_f,1,1}^* | l_{\Delta N_f,1,1} < r) = (P_{l_{\Delta N_f,1,1}}^* * P_l)(l_{\Delta N_f,1,1}^*) \quad (20)$$

and equivalent for row four. $P_{\Delta N_f,0}(\Delta N_f)$ can be calculated in the same way by interchanging all d_i and l_i and calculating l_s with $P_d(d)$ instead of $P_l(l)$.

We now calculate the difference between \mathbf{R}^{int} and \mathbf{R}^{test} . Afterward, we can add up the changes. When interpolating the test series, the RP (\mathbf{R}^{int}) regains the size of the reference RP (\mathbf{R}^{ref}), and every

anchor point is equal in both plots (Fig. 6),

$$R_{ij}^{\text{ref}} = R_{i/r, j/r}^{\text{test}} = R_{ij}^{\text{int}} \quad \text{for } i, j \in [0, r, 2r, \dots]. \quad (21)$$

If we make the assumption that the anchor diagonals have the same statistics for the length of 0- and 1-diagonal lines compared to all diagonals in the interpolated RP (\mathbf{R}^{int}), we only have to investigate the intervals between anchor points. For linear interpolation, there are four possibilities in \mathbf{R}^{int} . For the formal derivation, see Appendix B.

1. In intervals lying between two 1-anchor points, all points are 1; therefore, there is no 1-line starting point.
2. In intervals lying between two 0-anchor points, there is at most one 1-line in between; therefore, there is either one or zero 1-lines starting point.
3. In intervals lying between one 1- and one 0-anchor point, then there is one 1-0 transition and therefore, no start of a 1-line.
4. In intervals lying between one 0- and 1-anchor point, then there is one 0-1 transition and, therefore, one start of a 1-line.

Only the second point causes differences in the number of 1-lines between the plots because there is no start of a 1-line in \mathbf{R}^{test} , but there might be in \mathbf{R}^{int} . If there is a 1-line, we call this case a jump. The total number of jumps is N_j and can be directly measured by counting the number of intervals with jumps in \mathbf{R}^{int} . From this, we can follow that there is always a bigger or equal amount of 1-lines on anchor diagonals in the interpolated RP (\mathbf{R}^{int}) compared to the test RP (\mathbf{R}^{test}).

The estimated total difference ΔN_l of 1-lines lying on anchor diagonals between \mathbf{R}^{int} and \mathbf{R}^{ref} can now be calculated as the sum of, minus the number of intervals following a 1-anchor point times the mean number of 1-line starting-points in such an interval minus the number of intervals following a 0-anchor point times the mean number of 1-line starting points in such an interval plus the number of jumps,

$$\Delta N_l = -\frac{N_d}{r} RR \langle \Delta N_l \rangle_1 - \frac{N_d}{r} (1 - RR) \langle \Delta N_l \rangle_0 + N_j, \quad (22)$$

with N_d being the total number of points on the anchor diagonals, r is the interpolation ratio, and RR is the recurrence rate. $\frac{N_d}{r} RR$ is the number of intervals following a 1-anchor point, and $\frac{N_d}{r} (1 - RR)$ is the number of intervals following an 0-anchor point. The first two summands quantify the difference between \mathbf{R}^{test} and \mathbf{R}^{ref} , and $\langle \Delta N_l \rangle_1$ and $\langle \Delta N_l \rangle_0$ are calculated using Eqs. (9) and (10). The last summand quantifies the difference between \mathbf{R}^{int} and \mathbf{R}^{test} .

To get the estimated deviation of the average diagonal line length from this consideration, we first use Eq. (8) to write

$$\frac{L^{\text{int}}}{L^{\text{est}}} = \frac{N_l^{\text{int}} - \Delta N_l}{N_l^{\text{int}}} = 1 - \frac{\Delta N_l}{N_l^{\text{int}}}, \quad (23)$$

where L^{int} is the average line length measured in the interpolated RP (\mathbf{R}^{int}) and L^{est} is our estimation of L^{ref} measured in the reference RP (\mathbf{R}^{ref}). N_l^{int} is the number of 1-lines in the interpolated RP (\mathbf{R}^{int}) on the anchor diagonals, and ΔN_l is the estimated difference of the 1-lines between the reference and the interpolated RP on these diagonals.

After inserting Eq. (8) with $N_r = N_d \cdot RR$, which is the number of 1-points on the anchor diagonals, and Eq. (22), this can be written as

$$\frac{L^{\text{int}}}{L^{\text{est}}} = 1 + \left(\frac{\langle \Delta N_l \rangle_1}{r} + \frac{\langle \Delta N_l \rangle_0}{r} \cdot \left(\frac{1}{RR} - 1 \right) - \frac{N_j}{N_d \cdot RR} \right) L^{\text{int}}. \quad (24)$$

The total number of jumps N_j can also be rewritten as a probability n_j that a random interval contains a jump by dividing the number of jumps N_j by the number of intervals N_d/r ,

$$n_j = \frac{N_j \cdot r}{N_d}. \quad (25)$$

Inserted in Eq. (24), this gives the final result

$$\frac{L^{\text{int}}}{L^{\text{est}}} = 1 + \left(\frac{\langle \Delta N_l \rangle_1}{r} + \frac{\langle \Delta N_l \rangle_0}{r} \cdot \left(\frac{1}{RR} - 1 \right) - \frac{n_j}{r \cdot RR} \right) L^{\text{int}}. \quad (26)$$

This equation can be used to estimate the correction for the effects of the different sampling and the interpolation under the conditions that there is an integer interpolation ratio r and no offset between, the reference series and the test series. Furthermore, anchor and not anchor diagonals in the interpolated RP must have similar probability distributions for the 1- and 0-lines ($P_l(l)$ and $P_d(d)$), and the probabilities for consecutive lines have to be independent. In order to calculate our estimate L^{est} from a measured L^{int} from the interpolated RP, it is necessary to know the statistics for the 1- and 0-lines [$P_l(l)$ and $P_d(d)$] from the underlying dynamics. This might for a real-world example be accessible from a period in the data with a high sampling rate. The recurrence rate RR as well as the jumping probability n_j can be directly computed from the recurrence plot of the interpolated series. The recurrence rate is just the fraction of 1-points, and n_j is calculated by counting the intervals that contain a jump and dividing the results by the number of intervals.

In Sec. IV B, the use of this correction scheme is demonstrated on a simulated data processing example, where everything is known about the true underlying dynamics. Afterward, the scheme is used to correct for sampling and interpolation-induced biases in real-world data.

B. Results

To evaluate the correction scheme, $L^{\text{int}}/L^{\text{ref}}$ is calculated for the first autoregressive system and the Roessler system in the same way as before (Sec. III) and compared to the correction $L^{\text{int}}/L^{\text{est}}$, which is calculated with the described scheme. r being the interpolation ratio is varied, and while n_j and RR are obtained from the interpolated series and the corresponding recurrence plot (\mathbf{R}^{int}), $P_l(l)$ and $P_d(d)$ are obtained from the reference RP (\mathbf{R}^{ref}).

The estimated $L^{\text{int}}/L^{\text{est}}$ for the autoregressive process follows the increase in $L^{\text{int}}/L^{\text{ref}}$ with increasing interpolation ratio in the real data very closely if the interpolation ratio is smaller than 5 [Fig. 8(b)]. For higher values, there are some deviations caused by the small size of the test series. It has less than 100 data points for an interpolation ratio greater than 5. For the Roessler system, the correction only qualitatively captures the decrease, but with an underestimation of the effect [Fig. 8(a)].

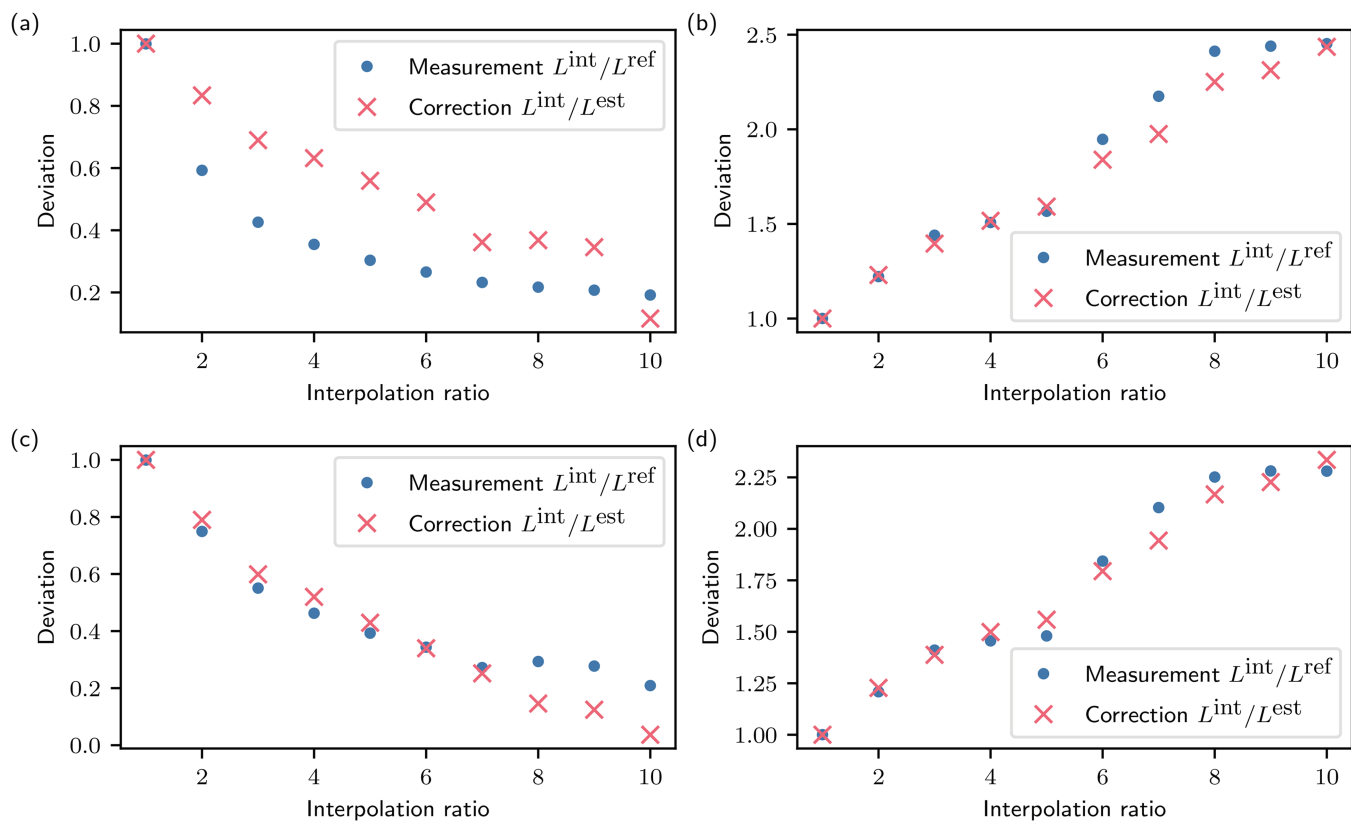


FIG. 8. Estimation and measured relative deviations of the L measure between the interpolated and reference series for the Rössler (top left) and autoregressive system (top right) (the same as in Fig. 4) for integer interpolation ratios. Equation (26) is used to calculate the correction. The bottom row shows the same systems, but the L -measure for the calculation of the change as well as for the calculation of the correction is obtained only from the anchor diagonals (Fig. 6).

This deviation is caused by the fact that the following assumptions are not true for the Rössler system. On the one hand, the statistics of the diagonal lines are very different when looking at all diagonals compared to the anchor diagonals. If only these diagonals are considered, the correction for the Rössler system captures the effect also quite well [Fig. 8(c)]. On the other hand, the assumption that the length of consecutive 1- and 0-lines is independent of each other is not true for a deterministic system, such as the Rössler system.

Using consideration from Sec. IV A, it is possible to understand the reasons for the differences between the systems. For the autoregressive process, the dominating effect is that there are fewer short 1- and 0-lines in \mathbf{R}^{int} compared to \mathbf{R}^{ref} ; therefore, the total number of lines is smaller and L^{int} is greater than L^{ref} . The dominating effect for the Rössler system is that there are more short 1-lines between two anchor points in \mathbf{R}^{int} than in \mathbf{R}^{ref} ; therefore, the total number of lines is greater and L^{int} is smaller than L^{ref} .

V. APPLICATION TO PALEOCLIMATE DATA

In this section, we use the insights from the theoretical considerations and simulations above to investigate real-world data.

We use the described correction scheme to reduce the influence of the interpolation and the different sampling times.

For this purpose, we use the data from a 290 m long composite core from Chew Bahir in southern Ethiopia. The core was created by combining the two ~ 280 m long parallel lacustrine sediment cores CHB14-2A and 2B, which were drilled within the Chew Bahir Drilling Project (CBDP) in 2014.³⁹ The aim of the CBDP was to establish a high-resolution environmental record spanning an important time interval of human evolution in eastern Africa. The potassium (K) concentration in the sediment is a proxy for the aridity in the region during the time of deposition.⁴⁰ It was determined by micro x-ray fluorescence (μ XRF) scanning, which was carried out with a spatial resolution of 5 mm. To attribute time points to the measured K concentrations, the age-depth model RRMarch2021⁴¹ is used, which follows a direct-dating approach and uses multiple chronometers, including AMS ^{14}C dating, optically stimulated luminescence (OSL) dating, argon-argon ($^{40}\text{Ar}/^{39}\text{Ar}$) dating, and tephrochronological data. Details about the age-depth model can be found in Roberts *et al.*⁴¹ In the following, we use this environmental record to show sampling and interpolation effects, which are not dependent on the accuracy of the age-depth model. A detailed RQA and corresponding interpretation has been performed in Trauth *et al.*⁴²

The age-depth model reveals a non-constant sedimentation rate; i.e., the data points in this time series are not sampled equally in time. We investigate the temporal change of the mean diagonal line length L using the sliding window approach with a window size of 3079 yr. Therefore, we get a measure of the temporal evolution of the predictability of climate. The mean sampling time changes between these windows. In the oldest part from 616 787 to 434 038 yrBP, the sampling time is around 15 yr, and from 434 038 to 148 149 yrBP, there is a sampling time of around 10 yr. In the newest part of the data from 148,149 yrBP until present, the sampling time changes a lot, from a minimum of around 5.36 to 47.27 yr [Fig. 9(a)]. The strong changes in the sampling time result from the increased number of age measurements available for more recent times. The long periods of uniform sampling times arise due to the limited availability of datable material within the cores during those time intervals.

To analyze the data, we use the measured potassium concentration together with the corresponding time points from the age-depth model as our time series. We interpolate this series linearly and

evaluate the interpolation function at time points with a fixed sampling time. To see the effect of a different choice of this sampling time, two time series are created, one with a sampling time of $dt = 5.36$ yr and one with $dt = 47.2$ yr. These sampling times correspond to the lowest and highest sampling times in the data. To avoid effects from the interpolation over long time periods with missing data (hiatuses), we exclude all points of the interpolated time series, where the temporal distance between the two neighboring measured data points exceeds 50 yr. For windows where more than 40% of the points are excluded, we do not calculate the L measure. This leads to some gaps in the L -series (see Fig. 9).

When comparing the L -curves calculated from the interpolated series with $dt = 5.36$ yr and $dt = 47.2$ yr, significant differences emerge. First, the first peak at approximately 10 000 yrBP is considerably stronger in the interpolated series with $dt = 5.36$ yr. Additionally, the amplitudes of the structures from 616 787 to 434 038 yrBP show an increase compared to the amplitudes of the structures between 434 038 to 148 149 yrBP [Figs. 9(b) and 9(c)]. The sampling time in the time intervals of increased amplitudes is lower than in the other parts, which leads to the hypothesis that the difference in the course of the data is due to the different interpolation. As we can see in Sec. III C, the deviations caused by an interpolation ratio smaller than 1 seem to be small compared to the effect of a larger interpolation ratio. This leads to the conjecture that the results from the interpolated series with $dt = 47.2$ are closer to the truth and the results of the series with $dt = 5.36$ yr are altered by the interpolation. To investigate this further, we apply our correction scheme to the L^{int} values. To use this method, the original distributions $P_l(l)$ and $P_d(d)$ have to be known. We assume some stationarity; i.e., these distributions should not change too much within the course of the data. We, therefore, use the histograms of 0- and 1-line lengths from the interval where the original dt is 5.36 yr. This is the case between 36 132 and 45 501 yrBP. RR , L^{int} , and n_j are calculated for every window. The correction method can only be used for integer ratios; therefore, the interpolation ratio is calculated for every window by dividing the mean sampling time in the data inside the window by the interpolation sampling time of 5.36 yr and rounded to the next integer to calculate the correction.

After applying the correction, the first peak is reduced in its height, and also, the changes of amplitudes in the older part of the data match the results of the interpolated series with $dt = 47.2$ yr [Fig. 9(d)]. This is an indication that the correction provides a close approximation of the true effect, which leads to the deviations in the L measure and helps to avoid misinterpretation due to the interpolation.

VI. CONCLUSION

Recurrence plots and recurrence quantification analysis are useful tools for the analysis of data from non-linear systems. They can give valuable insights into the changes in the dynamics and indicate critical regime changes. Applying these methods to paleoclimate data give additional and complementary valuable insights, which are usually not accessible with standard linear methods.

When using RPs for paleoclimate data, we have to take into account that the data from paleoclimate archives are usually not uniformly sampled in time. One commonly used method to tackle this

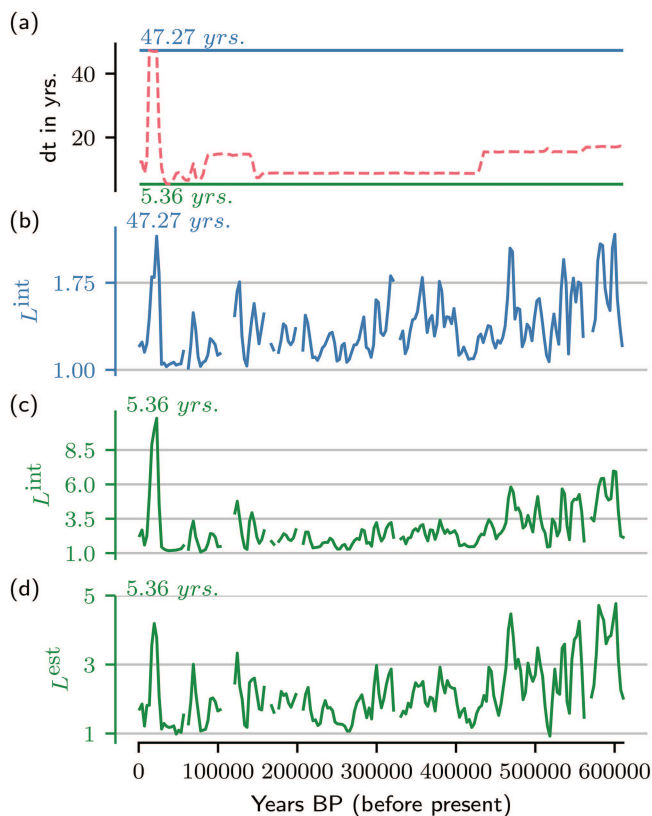


FIG. 9. (a) Mean sampling times for every window (red) and constant sampling times with which the interpolation function is evaluated (blue, green). The window step size is 3079 yr, and the window size is 6153 yr. (b) and (c) L for sliding windows, with $dt = 47.2$ yr (blue) and 5.36 yr (green). With fixed $\varepsilon = 250$. (d) Corrected data with a described scheme.

is to generate an interpolation function and evaluate it at equally spaced time points. This, however, can create some bias and lead to wrong conclusions.

In this work, we have demonstrated that, depending on the dynamics, the average diagonal recurrence line length L calculated from the RP of an interpolated signal can be greater or smaller compared to the true L derived from the RP of the raw signal with the same temporal resolution. We have shown this by comparing data with different sampling rates, which are interpolated to the time points of the reference time series.

For a Rössler system, L calculated from the interpolated series is smaller than the one calculated from the reference series for all considered interpolation methods. For autoregressive processes, on the other hand, L is bigger. Furthermore, we have explained that for an integer interpolation ratio, a possible offset can also change the result, and this is even true if the temporal resolution is not changed.

We identified three main reasons for the difference between the series with different sampling and interpolation: (1) distinct recurrence lines in the reference RP are merged in the interpolated recurrence plot, (2) short recurrence lines in the reference RP are missing in the interpolated one, and (3) there are short recurrence lines in the interpolated RP, which are not present in the reference one.

Using these insights, we developed a correction scheme and discussed its capabilities and limitations. For the autoregressive process, it can predict the difference very precisely, as long as the interpolation ratio is not too large and the downsampled data, therefore, not too short. For the Rössler system, on the other hand, it can capture the decrease but underestimates the effect up to a factor of two. Here, the main reason for this deviation seems to be that the correction is only true for the diagonal lines in the interpolated plot, which correspond to a diagonal in the downsampled one.

In the last part of our work, we have investigated paleoclimate data from a lake sediment core. First, we showed that the course of the L measure, when calculated in sliding windows, strongly depends on the choice of sampling time when evaluating the interpolation function. We then applied our proposed correction scheme to produce an approximation of the actual course.

This study shows that there are potentially big biases, when interpolating data, before applying the RP method. It also shows that the effect is strongly system dependent, and a simple correction might not be possible. The proposed correction scheme provides an intuition on which system shows which effects and also provides an approximate correction. The correction scheme is only valid for integer interpolation values without offset, which is rarely the case for real data. Nevertheless, this can give an approximation of the real effect. Further research is required to enhance our understanding of the impact of offset and noninteger interpolation ratios. Another limitation of this study is that the series with different temporal resolutions are obtained by downsampling, which is only similar to the measurement process if the measure yields the value of some quantity at one time point. In reality, measurements often do some kind of averaging over a finite period of time. To understand to what extent this changes the result of this study and how to change the correction scheme, further research is required.

ACKNOWLEDGMENTS

We thank Dr. Tobias Braun and Vanessa Skiba for fruitful discussions.

AUTHOR DECLARATIONS

Conflict of Interest

The authors have no conflicts to disclose.

Author Contributions

Nils Antary: Software (lead); Visualization (lead); Writing – original draft (lead); Writing – review & editing (equal). **Martin H. Trauth:** Writing – review & editing (equal). **Norbert Marwan:** Supervision (lead); Writing – review & editing (equal).

DATA AVAILABILITY

Code and data used for the research in this study are available in Zenodo at <https://doi.org/10.5281/zenodo.8123086>.

APPENDIX A: START LENGTH

When starting on a random 1-point in a recurrence plot, l_s is the remaining length on the 1-line, where this 1-point is a part of (see Fig. 7). To calculate the probability distribution of this length $P_{l_s}(l_s)$, we have to account for every combination of 1-line length together with the position of the point on this line, which creates a remaining line of the length l_s ; for example, if we start on 1-line with length 6 on the fourth position, the remaining length l_s is 2. The probability to hit a 1-line with a certain length is the probability to find such a line in the RP times the length of the line and this normalized,

$$P^*(l) = \frac{l \cdot P_1(l)}{\sum_{l=1}^{\infty} l \cdot P_1(l)}. \quad (\text{A1})$$

When hitting a 1-line of a certain length l , the probability that the remaining length is l_s is 0 if the length l is not longer than l_s . Otherwise, the probability is $1/l$. Summed over all possibilities, which are not zero, this gives in total

$$P_{l_s}(l_s) = \frac{\sum_{l=l_s+1}^{\infty} P_1(l)}{\sum_{l=1}^{\infty} l \cdot P_1(l)}. \quad (\text{A2})$$

APPENDIX B: RECURRENCE OF LINEAR INTERPOLATED POINTS

Here, we show the mathematical proof for the statements made about the 1-points on anchor diagonals.

- In intervals lying between two 1-anchor points, all points are 1.
 - The distance matrix for an interpolated time series is given by

$$D_{ij} = \|\vec{I}(t_i) - \vec{I}(t_j)\|, \quad (\text{B1})$$

where $\vec{I}(t)$ is the linear interpolation function. When looking at interpolation with integer r and without offset, we

TABLE II. Example of different intervals on the anchor diagonal of a recurrence plot and how they are changed when the underlying data are downsampled and linearly interpolated with an interpolation ratio of $r = 4$. The first and the last point are always the anchor points, as described in Sec. IV A.

Original	Downsampled	Interpolated	1-line-starts	Probability
11111	11	11111	0	$P(l_s > r)$
11000	10	11100	0	$P(l_s < r) \cdot P(l_s + d_1 > r \mid l_s < r)$
10011	11	11111	1	$P(l_s + d_1 < r) \cdot P(l_s + d_1 + l_1 > r \mid l_s + d_1 < r)$
10110	10	11100	1	$P(l_s + d_1 + l_1 < r) \cdot P(l_s + d_1 + l_1 + d_2 > r \mid l_s + d_1 + l_1 < r)$
10101	11	11111	2	$P(l_s + d_1 + l_1 + d_2 < r) \cdot P(l_s + d_1 + l_1 + d_2 + l_2 > r \mid l_s + d_1 + l_1 + d_2 < r)$

have anchor diagonals. At the anchor diagonals, the following equation holds:

$$t_i - t_j = z \cdot r \cdot dt_{\text{ref}} = z \cdot dt_{\text{test}} \quad z \in \mathbb{Z}, \quad (\text{B2})$$

with z being an integer and dt_{ref} and dt_{test} are the sampling times of the reference and test series. The interpolation function is a linear function for $t_k < t < t_{k+1}$, where t_k are the sampling points of the test series. Therefore, $\tilde{I}(t_i) - \tilde{I}(t_j)$ is the difference between two linear functions for $t_{k_i} < t_i < t_{k_i+1}$ and $t_{k_j} < t_j < t_{k_j+1}$. The second equation can be rewritten as $t_{k_j} < t_i - z \cdot dt_{\text{test}} < t_{k_j+1}$, which is the same as $t_{k_j} + z \cdot dt_{\text{test}} < t_i < t_{k_j+1} + z \cdot dt_{\text{test}}$, which is the same as the first conditions. This shows that the difference itself is a linear function between two anchor points. For the linear function, which goes through the points A and B , we can show that if $\|A\|$ and $\|B\|$ are smaller than ε , then all points in between have a norm smaller than ε . It follows that if our interpolation function is a linear function between two anchor points and both anchor points are recurrent and therefore, their norm is smaller than ε , all points on the interpolation function in between have a norm smaller than ε and are, therefore, also recurrent.

$$\begin{aligned} \|A\| < \varepsilon \quad \text{and} \quad \|B\| < \varepsilon, \\ C = A + z \cdot (B - A) \quad 0 \leq z \leq 1, \\ \|C\| = \|A + z \cdot (B - A)\| \\ &= \|(1 - z) \cdot A + z \cdot B\| \\ &\leq (1 - z) \cdot \|A\| + z \cdot \|B\| \\ &\leq \max(\|A\|, \|B\|) \\ &\leq \varepsilon. \end{aligned} \quad (\text{B3})$$

2. In intervals lying between two 0-anchor points, there is at most one 1-line in between.
 - Two 1-lines would violate the first point because there would be 0-points between 1-points.
3. In intervals lying between one 1- and one 0-anchor point, then there is one 1-0 transition.
 - There cannot be an additional 0-line before the last 1-point, as shown in the first point.
4. In intervals lying between one 0- and 1-anchor point, then there is one 0-1 transition.

- There cannot be an additional 0-line after the first 1-point, as shown in the first point.

APPENDIX C: EXAMPLE OF AN INTERVAL ON AN ANCHOR DIAGONAL

To illustrate changes to different intervals on an anchor diagonal, we show in Table II the effect of downsampling and interpolation. Furthermore, we give the probability of finding such an interval in the original recurrence plot.

APPENDIX D: INTERPOLATION EFFECT ON ROESSLER WITH A MAXIMUM NORM

To show that the effect seen for the Euler norm is also present when using the maximum norm, we performed the analysis from Fig. 4 (left) again, but using the maximum norm instead. Here, we also see that the L -measure obtained from the interpolated recurrence plot is decreased compared to the reference recurrence plot and the deviation is present for all interpolation methods and increases with the interpolation ratio. The linear and pchip interpolation leads to a similar deviation, which is qualitatively different from the deviation after using one of the spline interpolations (Fig. 10).

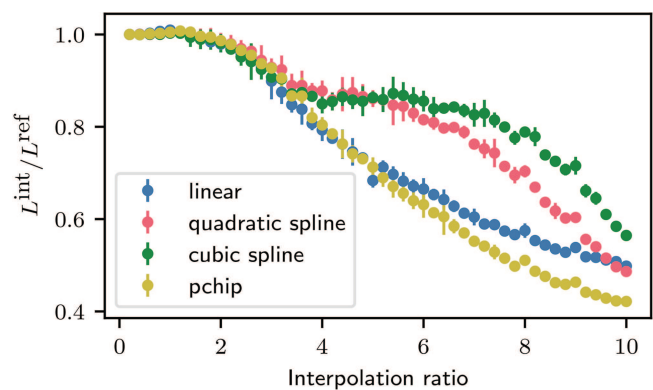


FIG. 10. Relative difference of L^{int} between interpolated and reference series. For a Roessler system, $L^{\text{ref}} = 8.77$, $\varepsilon = 4.45$, and the maximum norm is used.

REFERENCES

- ¹A. Wolf, J. B. Swift, H. L. Swinney, and J. A. Vastano, "Determining Lyapunov exponents from a time series," *Physica D* **16**, 285–317 (1985).
- ²X. Xu, J. Zhang, and M. Small, "Superfamily phenomena and motifs of networks induced from time series," *Proc. Natl. Acad. Sci. U.S.A.* **105**, 19601–19605 (2008).
- ³M. Martini, T. A. Kranz, T. Wagner, and K. Lehnertz, "Inferring directional interactions from transient signals with symbolic transfer entropy," *Phys. Rev. E* **83**, 011919 (2011).
- ⁴N. Marwan, M. C. Romano, M. Thiel, and J. Kurths, "Recurrence plots for the analysis of complex systems," *Phys. Rep.* **438**, 237–329 (2007).
- ⁵N. Marwan, "A historical review of recurrence plots," *Eur. Phys. J. Spec. Top.* **164**, 3–12 (2008).
- ⁶N. Marwan and K. H. Kraemer, "Trends in recurrence analysis of dynamical systems," *Eur. Phys. J. Spec. Top.* **232**, 5–27 (2023).
- ⁷N. Marwan, N. Wessel, U. Meyerfeldt, A. Schirdewan, and J. Kurths, "Recurrence plot based measures of complexity and its application to heart rate variability data," *Phys. Rev. E* **66**, 026702 (2002).
- ⁸U. Parlitz, S. Berg, S. Luther, A. Schirdewan, J. Kurths, and N. Wessel, "Classifying cardiac biosignals using ordinal pattern statistics and symbolic dynamics," *Comput. Biol. Med.* **42**, 319–327 (2012).
- ⁹N. Marwan, Y. Zou, N. Wessel, M. Riedl, and J. Kurths, "Estimating coupling directions in the cardio-respiratory system using recurrence properties," *Philos. Trans. R. Soc. A* **371**, 20110624 (2013).
- ¹⁰N. R. Lomb, "Least-squares frequency analysis of unequally spaced data," *Astrophys. Space Sci.* **39**, 447–462 (1976).
- ¹¹J. Ge, J. van Eyken, S. Mahadevan, C. DeWitt, S. R. Kane, R. Cohen, A. Vanden Heuvel, S. W. Fleming, P. Guo, G. W. Henry, D. P. Schneider, L. W. Ramsey, R. A. Wittenmyer, M. Endl, W. D. Cochran, E. B. Ford, E. L. Martin, G. Israelian, J. Valenti, and D. Montes, "The first extrasolar planet discovered with a new-generation high-throughput Doppler instrument," *Astrophys. J.* **648**, 683–695 (2006).
- ¹²S. F. M. Breitenbach, K. Rehfeld, B. Goswami, J. U. L. Baldini, H. E. Ridley, D. Kennett, K. Prufer, V. V. Aquino, Y. Asmerom, V. J. Polyak, H. Cheng, J. Kurths, and N. Marwan, "Constructing proxy-record age models (COPRA)," *Clim. Past* **8**, 1765–1779 (2012).
- ¹³I. Ozken, D. Eroglu, T. Stemler, N. Marwan, G. B. Bagci, and J. Kurths, "Transformation-cost time-series method for analyzing irregularly sampled data," *Phys. Rev. E* **91**, 062911 (2015).
- ¹⁴J. D. Scargle, "Studies in astronomical time series analysis. II. Statistical aspects of spectral analysis of unevenly spaced data," *Astrophys. J.* **263**, 835 (1982).
- ¹⁵P. Stoica and N. Sandgren, "Spectral analysis of irregularly-sampled data: Paralleling the regularly-sampled data approaches," *Digit. Signal Process.* **16**, 712–734 (2006).
- ¹⁶K. Rehfeld, N. Marwan, J. Heitzig, and J. Kurths, "Comparison of correlation analysis techniques for irregularly sampled time series," *Nonlinear Process. Geophys.* **18**, 389–404 (2011).
- ¹⁷I. Ozken, D. Eroglu, S. F. M. Breitenbach, N. Marwan, L. Tan, U. Tirnakli, and J. Kurths, "Recurrence plot analysis of irregularly sampled data," *Phys. Rev. E* **98**, 052215 (2018).
- ¹⁸R. Hébert, K. Rehfeld, and T. Laepple, "Comparing estimation techniques for temporal scaling in palaeoclimate time series," *Nonlinear Process. Geophys.* **28**, 311–328 (2021).
- ¹⁹N. Marwan, M. Thiel, and N. R. Nowaczyk, "Cross recurrence plot based synchronization of time series," *Nonlinear Process. Geophys.* **9**, 325–331 (2002).
- ²⁰T. Chelidze and T. Matcharashvili, "Dynamical patterns in seismology," in *Recurrence Quantification Analysis—Theory and Best Practices*, edited by C. L. Webber, Jr. and N. Marwan (Springer, Cham, 2015), pp. 291–334.
- ²¹S. Oberst, R. K. Niven, D. R. Lester, A. Ord, B. Hobbs, and N. P. Hoffmann, "Detection of unstable periodic orbits in mineralising geological systems," *Chaos* **28**, 085711 (2018).
- ²²A. Spiridonov, L. Balakauskas, R. Stankevic, G. Kluczynska, L. Gedminiene, and M. Stancikaite, "Holocene vegetation patterns in southern Lithuania indicate astronomical forcing on the millennial and centennial time scales," *Sci. Rep.* **9**, 14711 (2019).
- ²³A. Zaitouny, M. Small, J. Hill, I. Emelyanova, and M. B. Clennell, "Fast automatic detection of geological boundaries from multivariate log data using recurrence," *Comput. Geosci.* **135**, 104362 (2020).
- ²⁴S. Radzevicius, R. Stankevic, R. Budginas, A. Cichon-Pupienis, A. Venckute-Aleksiene, T. Meidla, L. Ainsaar, and A. Spiridonov, "Integrated stratigraphy of the Ludlow (Silurian) of the Baubliai-2 core (Western Lithuania) and the record of delta O-18 and delta C-13 climatically driven co-variability," *Newsl. Stratigr.* **56**, 75–88 (2023).
- ²⁵J. F. Donges, R. V. Donner, M. H. Trauth, N. Marwan, H. J. Schellnhuber, and J. Kurths, "Nonlinear detection of paleoclimate-variability transitions possibly related to human evolution," *Proc. Natl. Acad. Sci. U.S.A.* **108**, 20422–20427 (2011).
- ²⁶N. Marwan and J. Kurths, "Complex network based techniques to identify extreme events and (sudden) transitions in spatio-temporal systems," *Chaos* **25**, 097609 (2015).
- ²⁷D. Eroglu, F. H. McRobie, I. Ozken, T. Stemler, K.-H. Wyrwoll, S. F. M. Breitenbach, N. Marwan, and J. Kurths, "See-saw relationship of the Holocene East Asian-Australian summer monsoon," *Nat. Commun.* **7**, 12929 (2016).
- ²⁸M. H. Trauth, A. Asrat, W. Duesing, V. Foerster, K. H. Kraemer, N. Marwan, M. A. Maslin, and F. Schaebitz, "Classifying past climate change in the Chew Bahir basin, Southern Ethiopia, using recurrence quantification analysis," *Clim. Dyn.* **53**, 2557–2572 (2019).
- ²⁹T. Westerhold, N. Marwan, A. J. Drury, D. Liebrand, C. Agnini, E. Anagnostou, J. S. K. Barnet, S. M. Bohaty, D. De Vleeschouwer, F. Florindo, T. Frederichs, D. A. Hodell, A. E. Holbourn, D. Kroon, V. Laurentano, K. Littler, L. J. Lourens, M. Lyle, H. Pälike, U. Röhl, J. Tian, R. H. Wilkens, P. A. Wilson, and J. C. Zachos, "An astronomically dated record of Earth's climate and its predictability over the last 66 million years," *Science* **369**, 1383–1387 (2020).
- ³⁰N. H. Packard, J. P. Crutchfield, J. D. Farmer, and R. S. Shaw, "Geometry from a time series," *Phys. Rev. Lett.* **45**, 712–716 (1980).
- ³¹K. H. Kraemer, G. Datsis, J. Kurths, I. Z. Kiss, J. L. Ocampo-Espindola, and N. Marwan, "A unified and automated approach to attractor reconstruction," *New J. Phys.* **23**, 033017 (2021).
- ³²J.-P. Eckmann, S. Oliffson Kamphorst, and D. Ruelle, "Recurrence plots of dynamical systems," *Europhys. Lett.* **4**, 973–977 (1987).
- ³³J. P. Zbilut and C. L. Webber, Jr., "Embeddings and delays as derived from quantification of recurrence plots," *Phys. Lett. A* **171**, 199–203 (1992).
- ³⁴C. L. Webber, Jr. and J. P. Zbilut, "Dynamical assessment of physiological systems and states using recurrence plot strategies," *J. Appl. Physiol.* **76**, 965–973 (1994).
- ³⁵N. Marwan, J. F. Donges, R. V. Donner, and D. Eroglu, "Nonlinear time series analysis of palaeoclimate proxy records," *Quat. Sci. Rev.* **274**, 107245 (2021).
- ³⁶T. Braun, S. F. M. Breitenbach, V. Skiba, F. A. Lechleitner, E. E. Ray, L. M. Baldini, V. J. Polyak, J. U. L. Baldini, D. J. Kennett, K. M. Prufer, and N. Marwan, "Decline in seasonal predictability potentially destabilized Classic Maya societies," *Commun. Earth Environ.* **4**, 82 (2023).
- ³⁷K. H. Kraemer and N. Marwan, "Border effect corrections for diagonal line based recurrence quantification analysis measures," *Phys. Lett. A* **383**, 125977 (2019).
- ³⁸O. E. Rössler, "An equation for continuous chaos," *Phys. Lett. A* **57**, 397–398 (1976).
- ³⁹V. Foerster, A. Asrat, C. B. Ramsey, E. T. Brown, M. S. Chapot, A. Deino, W. Duesing, M. Grove, A. Hahn, A. Junginger, S. Kaboth-Bahr, C. S. Lane, S. Opitz, A. Noren, H. M. Roberts, M. Stockhecke, R. Tiedemann, C. M. Vidal, R. Vogelsang, A. S. Cohen, H. F. Lamb, F. Schaebitz, and M. H. Trauth, "Pleistocene climate variability in Eastern Africa influenced hominin evolution," *Nat. Geosci.* **15**, 805–811 (2022).
- ⁴⁰V. Foerster, A. Junginger, O. Langkamp, T. Gebru, A. Asrat, M. Umer, H. F. Lamb, V. Wennrich, J. Rethemeyer, N. Nowaczyk, M. H. Trauth, and F. Schaebitz, "Climatic change recorded in the sediments of the Chew Bahir basin, southern Ethiopia, during the last 45,000 years," *Quat. Int.* **274**, 25–37 (2012).
- ⁴¹H. M. Roberts, C. B. Ramsey, M. S. Chapot, A. L. Deino, C. S. Lane, C. Vidal, A. Asrat, A. Cohen, V. Foerster, H. F. Lamb, F. Schaebitz, M. H. Trauth, and F. A. Viehberg, "Using multiple chronometers to establish a long, directly-dated lacustrine record: Constraining > 600,000 years of environmental change at Chew Bahir, Ethiopia," *Quat. Sci. Rev.* **266**, 107025 (2021).
- ⁴²M. H. Trauth, A. Asrat, A. S. Cohen, W. Duesing, V. Foerster, S. Kaboth-Bahr, K. H. Kraemer, H. F. Lamb, N. Marwan, M. A. Maslin, and F. Schaebitz, "Recurring types of variability and transitions in the ~ 620 kyr record of climate change from the Chew Bahir basin, Southern Ethiopia," *Quat. Sci. Rev.* **266**, 106777 (2021).