Sampling rate-corrected analysis of irregularly sampled time series

Tobias Braun [*]
Potsdam Institute for Climate Impact Research (PIK), Member of the Leibniz Association, 14473 Potsdam, Germany

Cinthya N. Fernandez
Institute for Geology, Mineralogy and Geophysics Ruhr-Universität Bochum, 44801 Bochum, Germany

Deniz Eroglu
Faculty of Engineering and Natural Sciences, Kadir Has University, 34083 Istanbul, Turkey

Adam Hartland
Environmental Research Institute, School of Science, University of Waikato, Hamilton, Waikato 3240, New Zealand

Sebastian F. M. Breitenbach
Department of Geography and Environmental Sciences, Northumbria University, Newcastle upon Tyne NE1 8ST, United Kingdom

Norbert Marwan [†]
Potsdam Institute for Climate Impact Research (PIK), Member of the Leibniz Association, 14473 Potsdam, Germany

(Received 10 December 2021; accepted 4 February 2022; published 28 February 2022)

The analysis of irregularly sampled time series remains a challenging task requiring methods that account for continuous and abrupt changes of sampling resolution without introducing additional biases. The edit distance is an effective metric to quantitatively compare time series segments of unequal length by computing the cost of transforming one segment into the other. We show that transformation costs generally exhibit a nontrivial relationship with local sampling rate. If the sampling resolution undergoes strong variations, this effect impedes unbiased comparison between different time episodes. We study the impact of this effect on recurrence quantification analysis, a framework that is well suited for identifying regime shifts in nonlinear time series. A constrained randomization approach is put forward to correct for the biased recurrence quantification measures. This strategy involves the generation of a type of time series and time axis surrogates which we call sampling-rate-constrained (SRC) surrogates. We demonstrate the effectiveness of the proposed approach with a synthetic example and an irregularly sampled speleothem proxy record from Niue island in the central tropical Pacific. Application of the proposed correction scheme identifies a spurious transition that is solely imposed by an abrupt shift in sampling rate and uncovers periods of reduced seasonal rainfall predictability associated with enhanced El Niño-Southern Oscillation and tropical cyclone activity.

DOI: 10.1103/PhysRevE.105.024206

I. INTRODUCTION

The analysis of time series from complex systems calls for numerical methods that capture the most relevant features in the observed variability. At the same time, the impact of various frequently encountered data-related intricacies such as low signal-to-noise ratio, nonstationarity, and limited time series length must be accounted for. A major challenge is posed by irregular sampling, i.e., variations in the interval $\Delta_i = t_i - t_{i-1}$ between consecutive measurement times t_{i-1} and t_i. Irregular sampling is observed in many complex real-world systems. The underlying mechanisms that render the temporal sampling irregular may differ: sampling can be inherently irregular due to an additional process that controls the sampling interval (e.g., financial or cardiac time series [1,2]); a mixture of various external processes can result in "missing values," i.e., multiple interacting processes result in the nonavailability of measurements (e.g., sociological or psychological survey data [3]) or cause failures of the system (e.g., mechanical or electronical systems [4]); finally, the measurement process often results in irregularly sampled time series (e.g., astronomical [5] or geophysical systems [6]). Proxy time series obtained from palaeoclimate archives are a particularly challenging example since irregularity in the temporal sampling can itself contain valuable information on the processes of interest [7]. The growth rate of a stalagmite, for example, depends on variable environmental factors, including temperature in the cave and drip rate [8], among others. Since these factors and their variability are

[*]tobraun@pik-potsdam.de
[†]Also at University of Potsdam, Institute of Geosciences, 14473 Potsdam, Germany.

strongly coupled to the environmental conditions outside the cave, growth rate must be regarded as a dynamical indicator, for example, hydrological conditions, which in turn determine variations in the temporal sampling of the proxy time series.

Across many research communities, resampling based on interpolation techniques and imputation approaches are popular methods for making irregularly sampled time series compatible with standard time series analysis tools [9,10]. Artifacts and statistical biases caused by interpolation techniques are well known and may result in misinterpretation of the extracted time series properties, an issue further aggravated by the fact that biases introduced by interpolation may vary among different systems [11]. The robustness of results arising from different interpolation techniques for the same data set is rarely examined. For instance, linear interpolation will not compensate for the effect of lower variability during sparsely sampled episodes in a time series compared to more densely sampled periods. In fact, linear interpolation and mean imputation decrease variance to a hardly quantifiable, data-related degree [12]. Finally, more complex imputation models may account for such finite (sampling) size effects but may not represent the "natural" variability of a time series adequately. For data *not missing at random*, the assignment of a sufficient imputation model can be challenging and must account for nonstationarity in the underlying nonrandom effects (e.g., for the palaeoclimate example mentioned above). Similar biases are known from the problem of imbalanced data, i.e., given two populations that should be compared based on a statistical model, a majority class exists that contains significantly more samples than the minority class, and thus, oversampling techniques are applied to compensate for the resulting bias [13,14].

Geophysical time series frequently exhibit nonlinear features such as nonlinear oscillations and critical regime transitions, e.g., tipping points [15]. Dynamical system theory regards observations from such systems as embedded in a higher-dimensional phase space and offers a range of tools to quantify gradual or abrupt changes in these dynamics [16,17]. The power of these methods relies on their ability to uncover features that regular techniques, such as autocorrelations or variance estimation, fail to uncover [18]. Aiming for higher applicability of nonlinear time series analysis methods in the Earth sciences, irregular sampling approaches have been proposed [19–21]. One of these approaches is based on the idea of transforming subsequences of unequal lengths in a time series into each other and comparing the costs of these transformations for all subsequences [22]. More generally, the definition of a metric distance between states at different instances of time can entail dynamical information on the evolution of the phase space trajectory of the studied system. While standard metrics (such as Euclidean distance) fail to account for irregular sampling, the TrAnsformation-Cost Time-Series (TACTS) [23] includes the temporal information for distinct time series segments. The TACTS method is based on the edit distance measure, which was originally introduced to measure the similarity between marked point processes [22]. Similar approaches based on the edit or Levenshtein distance have been used in natural language processing [24] and metric analyses of point processes [25], among many others.

In this work, we focus on the application of the (m)Edit distance [25], which is a modified edit distance measure using a nonlinear transformation function instead of a scaler factor for measuring a cost operation. The modification helps to evaluate temporal patterns in sparse data sets such as paleoclimate proxies or extreme events. The time-sampling regularization by (m)Edit-distance preprocesses irregularly sampled time series for the computation of recurrence plots (RPs) [26]. The (m)Edit-distance approach can potentially be employed in any methodological framework that includes computation of a distance (or similarity) measure. The RP technique represents one particular application that has proven to be a powerful approach, tackling many of the fundamental problems in time series analysis, such as time series classification [27], the study of synchronization between multiple time series [28], and detection of regime transitions [29]. Recurrence quantification analysis (RQA) provides a means of quantifying the tendency of a time series to revisit previously visited states and has grown in its scope from basic predictability quantification towards more ambitious measures that, e.g., capture the multiscale nature of transitions [30–32]. The identification of shifts stands out as a particularly interesting application since critical transitions can often be linked to the vulnerability of the respective regional climate system towards external shocks or feedback mechanisms. The combination of the (m)Edit-distance approach and RPs offers a promising approach to identify regime transitions in irregularly sampled records, which may otherwise be impeded without an adequate technique designed to account for sampling variations [33–35]. In following this approach, special care must be taken if irregular sampling intervals undergo strong variations, i.e., where the process(es) that control the sampling rate are rendered nonstationary. In some applications, segments can be chosen such that they do not cover the same time period but the same number of values on average. Other applications require fixing a particular time period to be covered by each segment since this time period corresponds to the timescale under investigation, e.g., a year for seasonal time series. Even if such an approach is not motivated by the research question, splitting the time series into segments that correspond to nonequal time periods will result in mixing of timescales in the resulting distance matrix if the sampling rate is highly nonstationary. Here we focus on segments that cover equal time periods but varying numbers of values, referred to as *segment size*. We will show that in such cases, the resulting strong variations in segment size entail a nontrivial sampling bias of the (m)Editdistance.

We introduce the (m)Edit-distance methodology in Sec. II A followed by a short summary of recurrence analysis in Sec. II B. Section III illustrates the problem of strong variations in the sampling rate whereas model time series are studied to elucidate the sample size effects. A correction scheme based on constrained randomization is proposed in Sec. IV. In Sec. V we demonstrate the importance to correct for the identified sample-size dependence in an application to a palaeoclimate record from Niue island in the central Pacific where we identify variations in seasonal predictability. We conclude our findings in Sec. VI.

II. METHODOLOGY

A. The (m)Edit-distance measure

Many approaches in nonlinear time series analysis are based on some notion of a (dis)similarity measure. For deterministic systems, embedding the univariate time series into an m-dimensional phase space offers a multitude of quantitative approaches to analyze the variability of its trajectory [36]. Yet appropriate techniques to extract the embedding dimension and delay from empirical data are needed. These approaches can be cumbersome. In this work we focus on univariate time series wherein the most widespread dissimilarity measure between distinct segments $\mathcal{S}_a, \mathcal{S}_b$ is the Euclidean distance. It is a metric distance, i.e., its value is always positive $D(\mathcal{S}_a, \mathcal{S}_b) \geqslant 0$, it is symmetric $D(\mathcal{S}_a, \mathcal{S}_b) = D(\mathcal{S}_b, \mathcal{S}_a)$, and the triangle inequality holds $D(\mathcal{S}_a, \mathcal{S}_c) \leqslant D(\mathcal{S}_a, \mathcal{S}_b) + D(\mathcal{S}_b, \mathcal{S}_c)$. If the time series is characterized by missing values or the sampling interval Δ_i is irregular (e.g., due to irregularities in the measurement process), no straightforward application of Euclidean distance or comparable metrics is possible: dissimilarity of values at unequal timescales would be computed without accounting for their nonequality. Linear interpolation as a means of resampling the time series values onto a regular time axis is among the most popular approaches to regularize sampling [37]. Yet hardly controllable artifacts arise from linear interpolation, ranging from difficulties related to altered absolute timing to underestimation of variance or overestimation of persistence [11,38].

Originally proposed for natural language processing, the edit distance measure [39] is designed to compare sequences of variable length. Shifting and adding and deleting of strings were proposed as two elementary operations to quantify dissimilarities between words, an objective also pursued by other methods such as dynamic time warping [40]. The resulting costs are calculated by identifying a minimum cost path to transform one sequence into the other. Taking the next step towards an application to empirical time series, the edit distance was applied to point process data whereby cost parameters for the elementary operations remained arbitrary [22,41]. By equipping the technique with data-driven cost parameter estimates, it was then applied to irregularly sampled palaeoclimate time series [23]. A further modification [(m)Edit distance] with an application to extreme events was proposed to consider the saturation of shifting costs when a certain timescale τ, separating the two compared segments, is exceeded [25]. The main difference between applying the edit distance to series of events and spike trains and irregularly sampled time series is that for the latter, amplitudes of time series values must be considered. In the following, whenever no assumptions are made about the amplitudes of a signal, we refer to "events." The edit distance between two segments $\mathcal{S}_a, \mathcal{S}_b$ of an irregularly sampled time series is computed by minimizing the transformation costs by

$$D(\mathcal{S}_a, \mathcal{S}_b) = \min\left\{ \sum_{\alpha,\beta\in\mathcal{C}} \left[\underbrace{f_{\Lambda_0}(t(\alpha), t(\beta); \tau)}_{\text{shifting}} + \underbrace{\Lambda_\mathrm{k} \|L_a(\alpha) - L_b(\beta)\|}_{\text{amplitude change}} \right] + \underbrace{\Lambda_\mathrm{S}(|I| + |J| - 2|\mathcal{C}|)}_{\text{adding and deleting}} \right\} \tag{1}$$

with a norm $\|\cdot\|$ (e.g., the Euclidean norm), the αth and βth amplitudes $L_a(\alpha), L_b(\beta)$ of the segments $\mathcal{S}_a, \mathcal{S}_b$, and the cardinalities $|\cdot|$ of the sets I, J, and $\mathcal{C}$. While the latter are a set of indices of the time series values, $\mathcal{C}$ denotes the values that are shifted. $D(\mathcal{S}_a, \mathcal{S}_b)$ is a metric distance. The cost parameters Λ_0, Λ_k, and Λ_S need to be fixed prior to cost optimization. We choose the cost parameter for amplitudes changes Λ_k as suggested in [23]:

$$\Lambda_\mathrm{k} = \frac{M-1}{\sum_{i=1}^{M-1} \|x_i - x_{i+1}\|}. \tag{2}$$

The cost parameter Λ_S for deleting and adding has to be chosen such that deletions are neither "too cheap" nor "too expensive." For a set of time series values with a large temporal distance or very distinct amplitudes, a deletion and addition should be favorable, while a too low value of Λ_S will result in a transformation of sequences solely by deletion and adding operations even for very close time series values. We follow the scheme proposed in [33] by assuming normality for the distance values between all segments of the time series and optimize Λ_S within a specified range using a Kolmogorov-Smirnov (KS) test to ensure that the normality assumption holds as close as possible. Following the modification proposed in [25], costs associated with shifting of time instances between two time series values are controlled by the logistic function

$$f_{\Lambda_0}(t(\alpha), t(\beta); \tau) = \frac{\Lambda_0}{1 + e^{-[\|t_a(\alpha) - t_b(\beta)\| - \tau]}}, \tag{3}$$

where τ is the location parameter of the logistic function, reflecting a characteristic timescale that separates exponentially increasing from saturating or bounded exponentially increasing costs for shifting. We choose τ as the average sampling interval of the time series; $\tau = T/M$ with the total time period T and the number of samples M. Interpreting τ as a "temporal tolerance," this choice ensures that shifting exponentially fast becomes less favorable if time instances are separated by several standard deviations of the sampling interval distribution. Finally, a value for the maximum costs associated with shifting Λ_0 needs to be set. The ratio $\Lambda_\mathrm{K}/\Lambda_0$ reflects the relative importance of temporal and magnitudinal separation; in the limiting case $\Lambda_\mathrm{K}/\Lambda_0 \gg 1$, irregular sampling is no longer accounted for and the resulting distance between two segments solely reflects the norm $||L_a(\alpha) - L_b(\beta)||$ for all amplitudes $L_a(\alpha), L_b(\beta)$ of both segments $\mathcal{S}_a, \mathcal{S}_b$. In the opposite case $\Lambda_\mathrm{K}/\Lambda_0 \ll 1$, the time series can be regarded as a series of events since cost optimization is independent of their amplitudes. We choose $\Lambda_\mathrm{K} = \Lambda_0 = 1$. It must be stressed that this rate depends on the research question and the data under study.

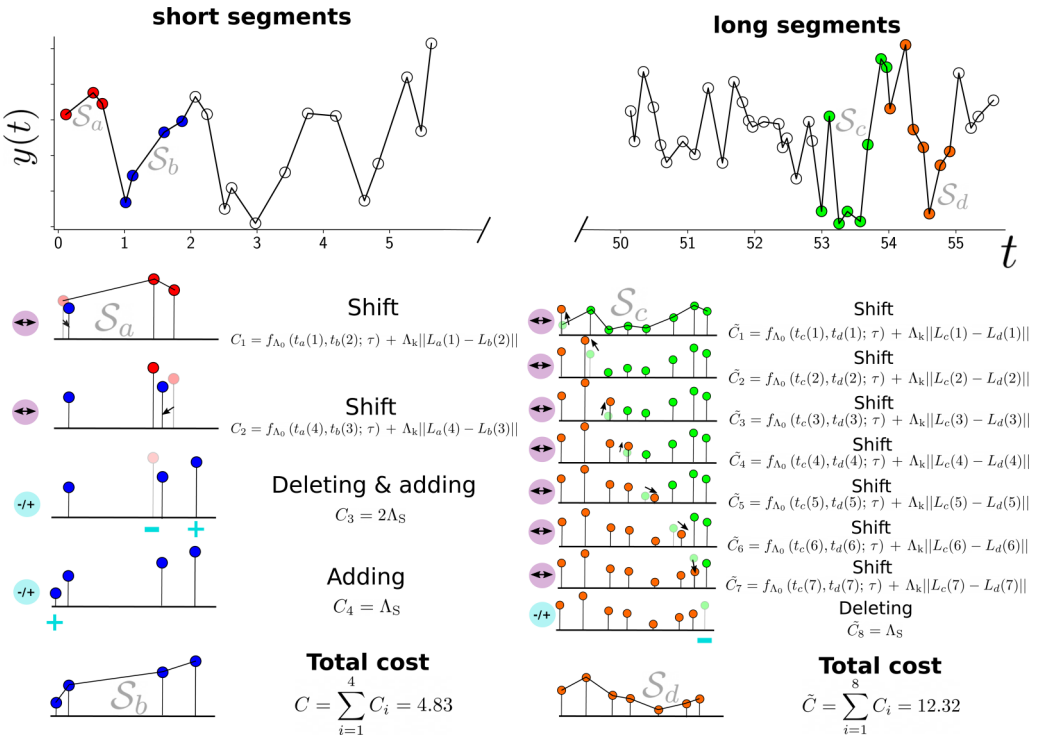


FIG. 1. Schematic illustration of how irregularly sampled segments of varying lengths are transformed with the (m)Edit-distance method. Two exemplary pairs of segments $\mathcal{S}_a, \mathcal{S}_b$ (a: red, blue) and $\mathcal{S}_c, \mathcal{S}_d$ (b: green, orange) of an irregularly sampled synthetic AR(1)-time series are displayed. Each row shows an operation applied to the respective segment (shift: purple, deletion or adding: cyan). Final costs C and $\tilde{C}$ result from a specific choice of cost parameters as described in Sec. II A. Note that the higher total cost in b showcases the dependence on segment length.

In the following, we discuss the finite-sample effects bias (m)Edit-distance values $D(\mathcal{S}_a, \mathcal{S}_b)$ and give a summary of the RP methodology. This facilitates the presentation of finite-sample effects discussed in Sec. III alongside an illustration of the (m)Edit-distance methodology (Fig. 1).

B. Recurrence analysis

The tendency to recur to previously visited states is a ubiquitous feature shared by time series from many different complex systems. Recurrence plots encode this information in a two-dimensional binary matrix, indicating a recurrence between two states $\vec{x}_i$ and $\vec{x}_j$ at times i and j if the respective states are similar with respect to a given norm $D(\vec{x}_i, \vec{x}_j)$ [42]:

$$R_{ij} = \begin{cases} 1 & \text{if } D(\vec{x}_i, \vec{x}_j) \leqslant \varepsilon \\ 0 & \text{if } D(\vec{x}_i, \vec{x}_j) > \varepsilon. \end{cases} \tag{4}$$

The norm $D(\vec{x}_i, \vec{x}_j)$ yields a symmetric, real-valued distance matrix $\mathbf{D}$ between states at all time instances i, j. By thresholding $\mathbf{D}$ with the vicinity threshold ε, a notion of similar and dissimilar states is implemented and defines the recurrence between each pair of states. The underlying idea is based on the Poincaré recurrence theorem that states the recurrence of a dynamical system's trajectory $\vec{x}(t)$ to an ε neighborhood of any perviously visited state after sufficiently long time [43]. For the main diagonal of the RP, it always holds that $R_{ij} \equiv 1$. If no phase space reconstruction is applied, states $\vec{x}_i$ and $\vec{x}_j$ correspond to time series amplitudes x_i and x_j. The threshold ε can be chosen based on different data-dependent criteria. In many applications, the recurrence rate is fixed to a certain percentage (e.g., 10% recurrences [44]) or set to a multiple of the standard deviation of the distance matrix $\mathbf{D}$ [45]. The geometric recurrence patterns encoded in a RP can be exploited to distinguish between stochastic and deterministic systems [26]; while a purely random white noise process will result in isolated dots in the recurrence matrix, time series from deterministic systems are known to yield diagonal line structures [26]. Long diagonal lines are characteristic for periodic systems; interrupted diagonal lines indicate chaotic dynamics. Recurrence quantification analysis (RQA), which evaluates the statistical properties of a RP, has proven a versatile tool for diverse real-world applications, such as time series classification [46], study of causal relations [47], or regime shift detection [48].

Recurrence analysis overcomes some of the flaws of other statistical analysis tools when applied to geophysical time series, such as the Lyapunov exponent or correlation dimension [49,50]. It is less sensitive to noise and can be applied to short time series. In combination with the (m)Edit-distance approach, first applications demonstrated its ability to detect regime transitions in palaeoclimate proxy records [23,51]. In order to compute a RP for irregularly sampled time series, $D(\vec{x}_i, \vec{x}_j)$ in Eq. (4) is identified with the modified edit distance from Eq. (1). In contrast to regular computation of metric distances, segments of the time series are required to obtain a distance value between two states. Generally speaking, segment size should be chosen sufficiently small to ensure that no

aliasing effects arise due to interference between the segment width and the characteristic timescale of a time series (e.g., characteristic period of a periodic time series). For some applications the segments can be chosen such that all are equally sized $|\mathcal{S}_a| = |\mathcal{S}_b| = \cdots = N$. If this is not possible, the variance of segment widths can still be minimized and for each pair of segments with differing widths; deletion and adding operations will contribute to the resulting transformation cost. If time series are short, we can allow for an overlap between segments, although caution is advised since this introduces a serial dependence in the resulting edit distances of overlapping segments and violates the normality assumption used in the estimation of Λ_S. Here we focus on the most general case of unequal segment sizes. Apart from cases where segment size deviations can hardly be minimized, this is relevant in some real-world applications where we are interested in the recurrences between segments that correspond to a particular timescale, or where sampling rate is highly nonstationary and selecting a constant segment size would result in mixing of distinct timescales. The application to palaeoclimate data (Sec. V) will illustrate such a case. There the focus lies on the comparison of seasonal sequences in an irregularly sampled proxy time series.

Predictability is a feature of time series that can help to identify and classify different dynamical regimes in the evolution of the studied system. Since the lengths of diagonal lines in a RP reflect the predictability of a system, the number of diagonal lines which exceed a specified minimum line length $l_{\min}$ can be used as a predictability measure:

$$\mathrm{DET} = \frac{\sum_{l=l_{\min}}^{N} P(l)}{\sum_{l=1}^{N} P(l)} \tag{5}$$

with the number $P(l)$ lines of length l. Determinism (DET) can be linked to the correlation dimension of a dynamical system [52] and has successfully been used in diverse empirical analyses [33,34,48] to detect transitions between regimes of varying predictability. We use DET as a recurrence quantifier to test the impact of the sampling-based correction scheme introduced below.

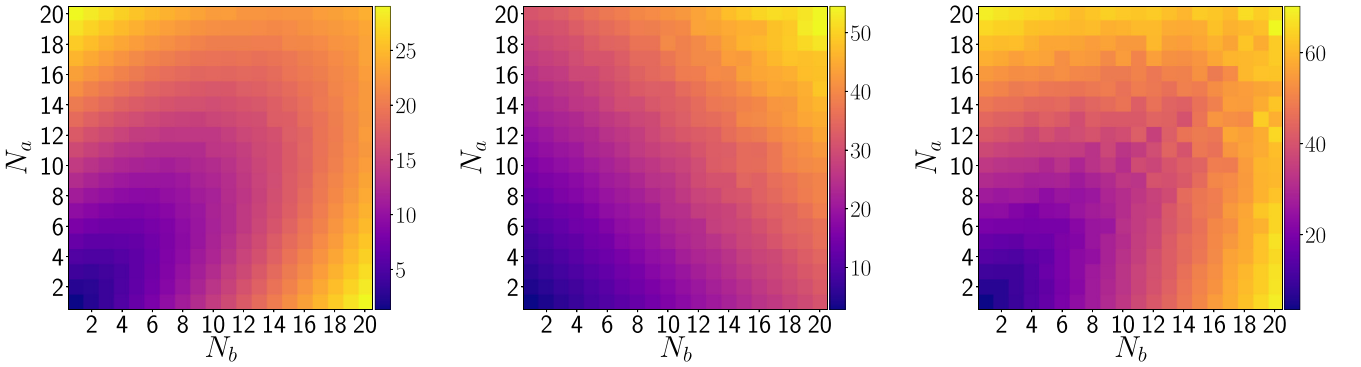
III. SEGMENT SIZE DEPENDENCE

Finite-sample effects are known to entail statistical biases in various time series analysis methods. Linear or spline interpolation is often employed as a preprocessing technique to enable the application of standard time series analysis tools to irregularly sampled time series. Interpolation techniques do not account for basic finite-sample biases. For instance, statistical location and scale measures (such as the median or volatility indicators) are known to be biased for small sample sizes [53,54]. Given two segments $\mathcal{S}_a, \mathcal{S}_b$ with $|\mathcal{S}_a| \gg |\mathcal{S}_b|$, estimating their variance (e.g., as a volatility indicator or in order to compute a continuous wavelet spectrum) can result in underestimation of the variance for the shorter segment. Similarly, persistence estimators are generally biased due to finite-sample effects, even for Markovian stationary stochastic processes [55]. Whenever a sliding-window analysis for nonstationary, irregularly sampled time series is carried out, variations in the sampling rate will inevitably result in a mixture between the actual variability of the statistical indicator and purely sampling-related variations. As interpolation techniques are usually limited to resampling values such that sampling intervals are equal, this effect is not compensated. Similar intricacies need to be considered in short time series, e.g., when computing correlations between multiple time series (of varying length) [56].

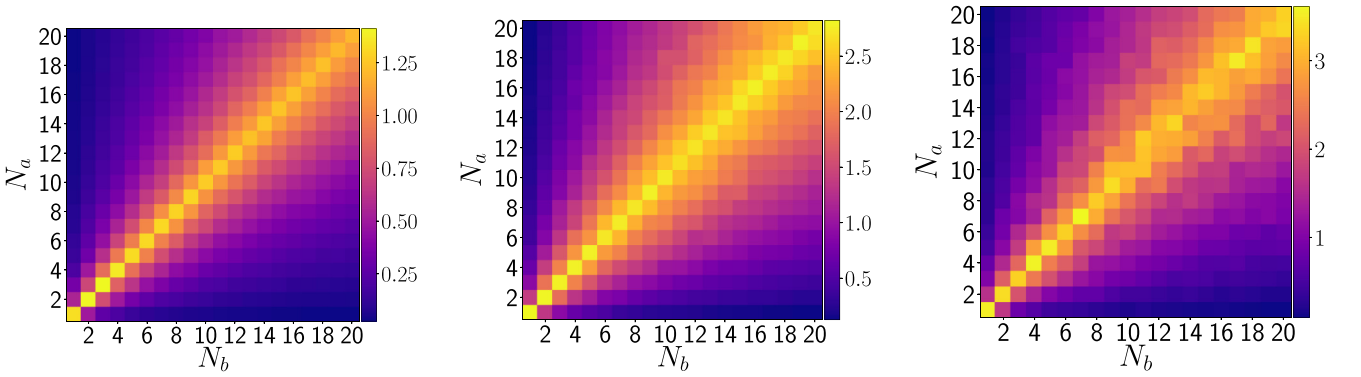
While not designed to compensate such effects, the (m)Edit-distance methodology does not introduce any known additional biases. The computation of transformation costs is demonstrated with two exemplary pairs of segments $\mathcal{S}_a, \mathcal{S}_b$ and $\mathcal{S}_c, \mathcal{S}_d$ (Fig. 1). The segments $\mathcal{S}_a, \mathcal{S}_b$ all display distinct operations for transforming a segment into another: in the first step, a shift of amplitude and time are applied to transform the time instance $t_a(1)$ and amplitude $L_a(1)$ of the first segment into time instance $t_b(2)$ and amplitude $L_b(2)$ of the second segment. The cost C_1 associated with this operation is the sum of shifting both time and amplitude. After shifting the third value of $\mathcal{S}_a$ to match the third value of $\mathcal{S}_b$, both a deletion and an adding operation are performed in step 3 with twice the cost Λ_S for a adding and deleting operation. The same transformation could have been achieved with an additional shifting operation. The preferred operation is determined by the particular choice of cost parameters. As $|\mathcal{S}_a| = 3$ and $|\mathcal{S}_b| = 4$, the first value of $\mathcal{S}_b$ is added in step 4. The resulting cost is the sum of all costs for each step. While different transformation paths are possible, the algorithmic implementation ensures that C is minimized with respect to all possible combinations. Another example is displayed in the right column of Fig. 1. The setup differs in that the indicated segments $\mathcal{S}_c, \mathcal{S}_d$ are longer than $\mathcal{S}_a, \mathcal{S}_b$ ($|\mathcal{S}_c| = 8, |\mathcal{S}_d| = 7$). Despite a similar set of transformations, the resulting costs $\tilde{C}$ are significantly higher for the exemplary choice of parameters.

A systematic derivation of transformation costs on segment size or sampling rate for exponentially distributed sampling intervals is given in Appendix A. Note that the identified effect is not due to an immanent misconception in the edit distance computation. It solely arises from the fact that the edit distance is applied in a setting where the time axis is not only irregular but undergoes significant variations in its sampling rate. In particular, abrupt transitions in the sampling rate between a time period T_1 with low sampling rate λ_1 and T_2 with high sampling rate λ_2 will imprint a nontrivial λ_1, λ_2-dependence on the transformation cost $D(\mathcal{S}_a\mathcal{S}_b)$ between any two segments. In a recurrence analysis of time series, the focus lies on the similarity of states based on the amplitudes of the time series. Hence, we argue that the identified dependencies counteract the goal of recurrence analysis of irregularly sampled time series and thus need to be corrected such that recurrence quantification measures reflect the dynamical behavior of the underlying system rather than mere shifts in the sampling rate.

We numerically examine the dependence of transformation costs between segments $\mathcal{S}_a, \mathcal{S}_b$ on their sizes N_a, N_b for simple synthetic time series. We test irregularly sampled time series from three different model systems: uncorrelated uniform noise, an AR(1)-process ($\tau = 5$), and a sinusoidal ($\nu = 1/25$) with superimposed low-amplitude white noise. Segments of specified sizes from each of these systems are drawn to compute segment size-specific costs.



(a) Cost matrices $C(N_a, N_b)$ (including basic deletions, total costs).



(b) Cost matrices $\tilde{C}_{\text{shift}}(N_a, N_b)$ (excluding basic deletions, average costs per operation).

FIG. 2. Cost matrices $C(N_a, N_b)$ for the transformation of segments with different lengths, including basic deletions (a) and excluding basic deletions (b). Costs are shown for uncorrelated uniform-distributed noise (left), an AR(1)-process (center), and a sinusoidal with superimposed white noise (right). Sampling intervals are γ-distributed.

Irregular time axes are generated from a $\gamma(\Delta; k, \Theta)$-distribution with scale Θ and shape $k = \sqrt{2/\Gamma}$, where Γ denotes the skewness of the distribution. This choice is motivated by the observation that sampling intervals in palaeoclimate proxy time series are often γ-rather than exponentially distributed. For each system, we generate a "superpopulation" ($K = 100$) of time series and time axes. Fixing a different skewness Γ of the γ-distribution of each of the time axes between $\Gamma \in [1, 8]$ ensures that for $T = 10\,000$, segment sizes range between $N \in [1, 20]$. The (m)Edit-distance is used [Eq. (1)], and deletions are included as a competing operation to shifting. The optimal Λ_S is estimated for each system according to the procedure outlined in Sec. II A: the KS statistic is minimized for each systems, yielding $\Lambda_S^{(\text{unif})} = 1.5$, $\Lambda_S^{(\text{AR1})} = 1.5$, $\Lambda_S^{(\text{sin})} = 3.5$.

Figure 2(a) displays the obtained transformation costs in the cost matrices $C(N_a, N_b)$ and $\tilde{C}_{\text{shift}}(N_a, N_b)$ after averaging over $K = 100$ different realizations. Regardless of the irregularity of the time axis and the respective system, a tendency of increasing total costs for larger segment sizes is observed (upper row). For the AR(1)-system, this increase is slower for fixed N_b and increasing N_a composed to the uncorrelated noise and the sinusoidal examples. More generally, the rate of increase differs between the considered systems but follows the same trend. In total, $|N_a - N_b|$ "basic deletions" (or adding operations) need to be carried out for each pair of segments with $N_a \neq N_b$. If costs for these basic deletions are subtracted and computed per shifting step, a similar dependency on N_a, N_b as observed in Fig. 7(c) for the more simple case can be observed in the cost matrices $\tilde{C}_{\text{shift}}(N_a, N_b)$ in Fig. 2(b): the cost of an average shift from a segment with $N = N_a$ increases towards $N_b = N_a$ and decays if segment size increases further. Consequently, the leading effect results from the basic deletions that are directly linked to the difference in segment sizes $|N_b - N_a|$. Yet transformation costs still depend on segment size even after aligning both segment sizes by means of basic deletions; this effect likely results from having a higher probability of finding closely spaced values on the time axis as the sampling rate of one segment increases, yielding an increasing trend for average costs per operation [in Fig. 2].

IV. SAMPLING RATE CONSTRAINED SURROGATES

Irregularly sampled time series with constant sampling rate can be studied with the (m)Edit-distance to obtain dissimilarity estimates between different time series segments. The resulting distance matrix can be used to perform a recurrence analysis. Moreover, other analysis techniques such as complex networks, clustering, or correlation analysis are based on (dis)similarity measures and could use the (m)Edit distance as a metric to account for irregular sampling or to characterize event-like data. In Sec. III we showed that in

case of a nonconstant sampling rate, an estimation of the (m)Edit-distance matrix is biased by significant differences in the segment sizes.

In the following, we propose a numerical correction technique for recurrence analysis. We generate an ensemble of time series and time axis surrogates that reproduces the sampling properties of the real irregularly sampled time series. This surrogate ensemble is used for bias correction of recurrence quantification measures, exemplified by the determinism DET.

A. Constrained randomization

When studying a system's dynamics with time series analysis tools, a null hypothesis is formulated which can be be tested. In case of recurrence analysis, this hypothesis could for example be nonstationarity of a dynamical property of the system (predictability, serial or cross-dependence, etc.) expressed by a particular recurrence quantification measure. In the used example, the null hypothesis tests whether the observed dynamics could be solely caused by variations in the sampling rate.

Parametric hypothesis testing for time series analysis often poses severe constraints on the statistical properties of the underlying probability distribution, e.g., normality. Surrogate tests represent a nonparametric and flexible method to test for a range of properties in a system, including nonlinearity or periodicity, among others [57–59]. Time series surrogates are altered copies of a real, underlying time series that preserve only a specified set of properties of the real time series. The general technique to generate surrogate realizations of a time series is constrained randomization [60]. After defining a set of constraints that state which properties of the real time series should be preserved, the time series is randomized such that these constraints are still fulfilled. Here randomization will be carried out on the sampling interval Δ_i with the constraint that for each segment $\mathcal{S}_i$ of the real time series, segment size N_i is preserved. This is achieved by drawing sampling intervals Δ_i (with replacement) from the empirical sampling interval distribution $p(\Delta, \lambda(t))$. For a given segment S_i with size N_i, N_i sampling intervals are drawn from $p(\Delta, \lambda(t))$ and cumulated to generate a surrogate realization of the particular time axis segment:

$$\tilde{t}_{\mathcal{S}_i}^{(0)} = t_{\mathcal{S}_i}^{(0)}, \quad \tilde{t}_{\mathcal{S}_i}^{(j+1)} = \tilde{t}_{\mathcal{S}_i}^{(0)} + \sum_{m=0}^{j} \Delta_i^{(m)}. \tag{6}$$

Let w be the time period covered by each segment. For any randomly sampled set of sampling intervals, the constraint of preserved segment size requires that

$$\tilde{t}_{\mathcal{S}_i}^{(N_i)} \overset{!}{\leqslant} w, \tag{7}$$

otherwise the random sampling of sampling intervals Δ_i has to be repeated. If the distribution of segment sizes is short-tailed, i.e., no segments with size $N \gg \mathbb{E}[k]$ exist, this simple randomization procedure converges rapidly for each segment. If segments of relatively large size are present, which is likely the case for nonstationary sampling rates, only a small subset of sampling intervals from the left tail of $p(\Delta, \lambda(t))$ will fulfill the condition (7). In order to ensure convergence of the algorithm for large segments, a weight function can be introduced for all sampling intervals to increase the likelihood of drawing short sampling intervals when a segment with large size is generated. We suggest the use of β-distributed weights ω:

$$\omega(X; \alpha, \beta) = \frac{1}{B(\alpha, \beta)} x^{\alpha-1}(1-x)^{\beta-1} \tag{8}$$

with the β function $B(\alpha, \beta)$. This choice is motivated by the fact that for $\alpha = \beta = 1$, $\omega(X; \alpha, \beta)$ becomes a uniform distribution. In our application, we choose $\alpha = \beta = 1$ when the first iteration of sampling N_i sampling intervals Δ_i is carried out. The population of sampling intervals is ordered from shortest to largest and each $x_i \leftrightarrow \Delta_i$ is assigned a β-distributed weight ω_i, i.e., for the first iteration, every sampling interval is drawn with equal probability. If the iteration fails ($\tilde{t}_{\mathcal{S}_i}^{(N_i)} > w$), α is increased by a small number $\Delta\alpha$, reshaping the beta distribution and increasing the probability of drawing small sampling intervals. Thus, we perform a weighted sampling from the empirical distribution $p(\Delta, \lambda(t))$ of sampling intervals with β-distributed weights. In the lth iteration, we use $\omega(X; \alpha_l, \beta = 1), \alpha_l = 1 + l\Delta\alpha$ as the weight function for each segment. Finally, we can identify an amplitude difference Δy_i of the time series with each sampling interval Δ_i. This correspondence is exploited by also drawing the respective amplitude difference for each drawn sampling interval. After the procedure is finalized and a surrogate has been generated, amplitude differences are cumulated, and that yields both a time axis and time series surrogate. Both are denoted as sampling-rate-constrained (SRC) surrogates. The full randomization procedure thus preserves segment sizes in the correct temporal order and by definition approximately reproduces the distribution of amplitude differences and sampling intervals.

It also preserves the correspondence between sampling intervals and amplitude differences, ensuring that if periods with high local sampling rate entail larger variance or strong amplitude changes in a real time series, this property is also included in the SRC surrogates. The full procedure is outlined in Fig. 3 for an exemplary time series. Other randomization schemes are conceivable, e.g., varying the sampling weights after drawing each single sampling interval based on the size of the latter, or stratified randomization, i.e., performing the randomization differently for strata that correspond to the different segment sizes. However, the proposed scheme has proven to be effective within the scope of this work.

With the presented scheme of generating SRC surrogates, an ensemble of surrogates can be generated and (m)Edit-distance matrices $\mathbf{D}$ computed for each SRC surrogate. Any measure that is based on $\mathbf{D}$ can consequently be computed for each surrogate separately, yielding a distribution that can be used for testing the null hypothesis formulated above based on the desired α-confidence level.

B. Recurrence analysis of an AR(1) process

In the example below, the proposed correction scheme is applied to an irregularly sampled AR(1)-process [Fig. 4(a)]. We consider an autocorrelation increasing with time, visible by autocorrelation time τ [Fig. 4(b)]. A recurrence analysis is

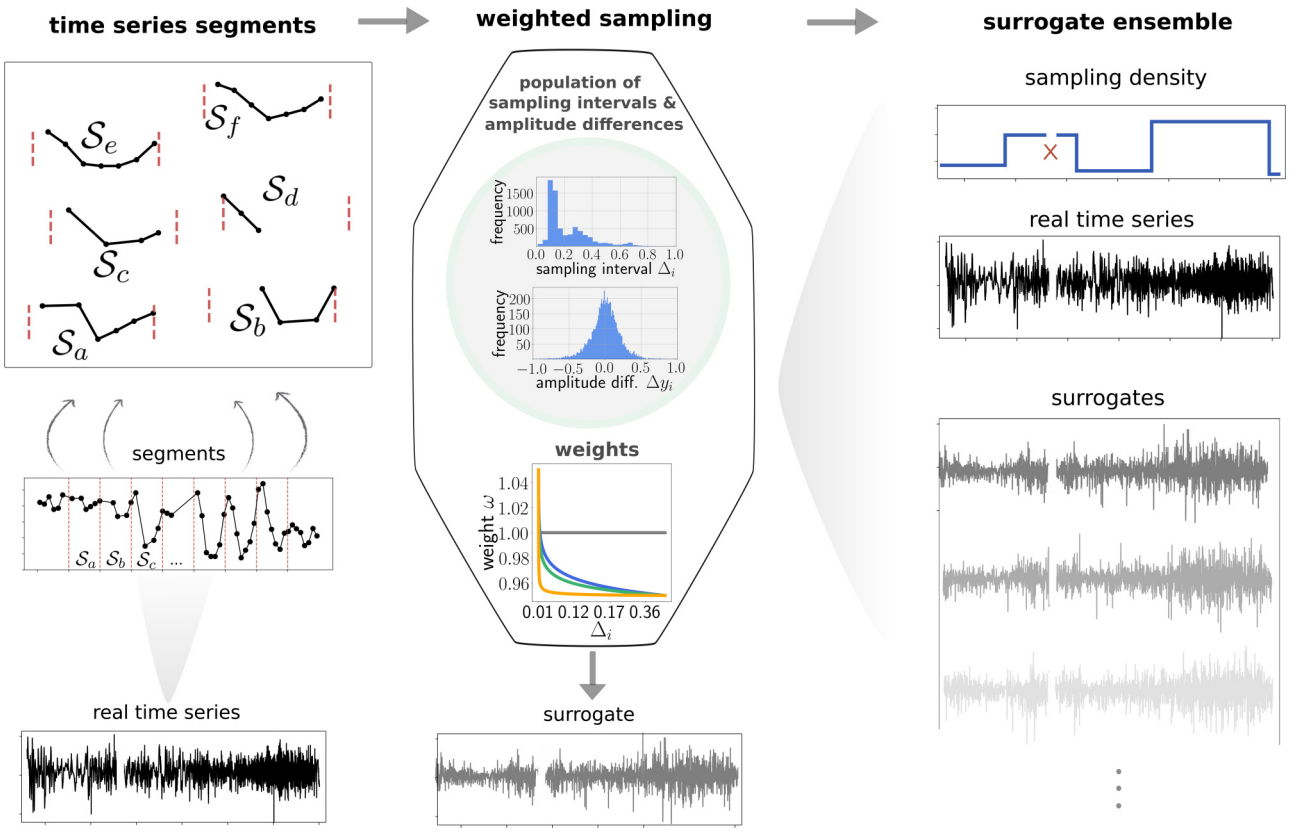


FIG. 3. Schematic illustration of the constrained randomization procedure that generates SRC surrogates for an exemplary irregularly sampled time series with nonstationary sampling rate. The left column shows the segmentation of the time series into segments of constant time period s but of variable size N_i. The center column illustrates the weighted sampling of sampling intervals and amplitude differences. Each sampling interval is assigned a β-distributed weight whereby the α parameter of the weight distribution is increased with each lth failed iteration to favor short sampling intervals. The resulting surrogates preserve the empirical distributions and segment sizes. Since amplitude differences are sampled jointly with the respective sampling intervals, increased volatility simply due to a higher local sampling rate is reproduced by the SRC surrogates.

used to characterize the predictability of the time series in a sliding window analysis. Predictability is computed by means of determinism, DET, as defined in Eq. (5). In particular, we study how an abrupt shift of the sampling rate (represented by the skewness of γ-distributed sampling intervals) affects DET and if a continuous increase of predictability can be recovered despite this shift by using the proposed SRC-surrogate method. The shift appears at $t = 1250$ [visible by variation of the segment size; Fig. 4(b)].

We expect DET to reproduce the linear increase in autocorrelation time, because increased serial dependence implies longer and more diagonal lines in the RP. For the computation of the (m)Edit-distance measure, segments are picked such that each covers a constant time interval of $w = 1$ which could correspond to a year in a real-world application. 200 SRC-surrogates are generated (see Appendix B) with $\alpha_0 = 1$, $\beta = 1$ and a step size for the shape parameter α of the beta distribution $\Delta\alpha = 0.15$. We set an upper limit of $N_{\text{it}}^{(\max)} = 1000$ for the number of iterations in the generation of each segment which is never exceeded in the performed simulations. The deletion and adding cost parameter Λ_S is estimated separately for the real time series and the surrogate realizations, yielding $\Lambda_S^{(\text{real})} = 5.3$ and $\Lambda_S^{(\text{SRC})} = 2.6$. Recurrence plots are computed on sliding windows of size $s = 200\Delta$ with 75% overlap (time series length: $T = 5000$). We fix a recurrence rate of 15% and do not apply any time-delay embedding. For each window, two DET values are obtained [Fig. 4(c)]: the DET value of the real time series and the $\alpha(= 95\%)$ confidence level of DET values calculated from the SRC-surrogate ensemble. The DET measure indicates a spurious transition of predictability induced by the abrupt shift in sampling rate [Fig. 4(c), gray shading]. Both the real time series and the surrogate ensemble indicate this shift, demonstrating that the proposed SRC-surrogates effectively reproduce the sampling bias. The SRC-based correction is applied to DET values by dividing the real DET-series by the 95% confidence level for each window [Fig. 4(d)]. The resulting predictability estimates reproduce the expected linear increase in serial dependence while eliminating the spurious shift due to the jump of the sampling rate.

V. REAL-WORLD APPLICATION: RAINFALL SEASONALITY IN THE CENTRAL PACIFIC

Many real-world proxy time series are characterized by irregular sampling or missing data and stationarity of the underlying process that controls the sampling rate cannot be guaranteed. This perspective even goes beyond uneven time axes

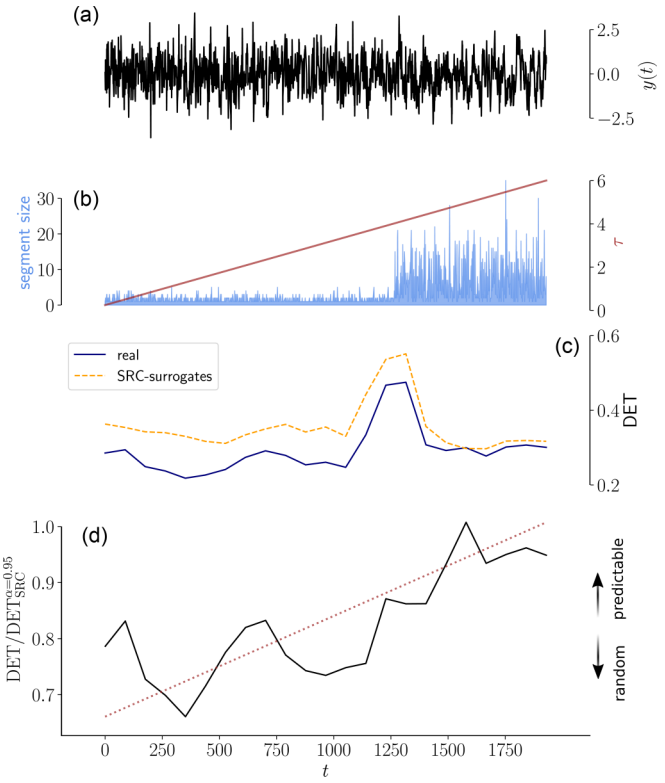


FIG. 4. Application of SRC-surrogate correction method to (a) an irregularly sampled AR(1)-process with (b) nonstationary sampling rate (blue) and linearly increasing autocorrelation time (red). Gray shading indicates the abrupt shift in sampling rate. (c) A sliding window RQA using determinism (DET) as a predictability measure is carried out. Real DET values are displayed in dark blue, the 95% confidence level computed from 200 SRC surrogates is shown in yellow. (d) The ratio $\mathrm{DET}_{\mathrm{real}}$ by $\mathrm{DET}_{\mathrm{surr}}$ provides a sampling-bias corrected predictability measure that reproduces the linear increase in serial dependence.

as for some systems, it might be desirable to apply an adaptive windowing in order to obtain segments with segment sizes depending on specific parameters of the system. For instance, when analyzing cardiac time series it might be reasonable to choose the segment size adaptively to capture one heart-beat cycle within each segment. The length of every cycle is controlled by a variety of other physical, nonstationary parameters. Below we focus on an irregularly sampled palaeoclimate proxy time series with a nonstationary temporal sampling rate. We demonstrate the effectiveness of the proposed approach by carrying out a sliding window recurrence analysis.

The palaeoclimate record analyzed here is a seasonally resolved stalagmite proxy record from Niue island in the southwestern Pacific (19°S, 169°W). It covers 1000 years in the mid-Holocene [6.4–5.4 thousand years before present (ka BP)]. Niue island has a tropical climate, receiving an average of 2000 mm of precipitation annually with a pronounced wet season from November to April. Rainfall is most strongly controlled by seasonal displacement of the South Pacific Convergence Zone but also reacts sensitively to atmospheric circulation changes associated with the El Niño-Southern Oscillation. Here we analyze seasonal rainfall variability on Niue recorded in grayscale changes that

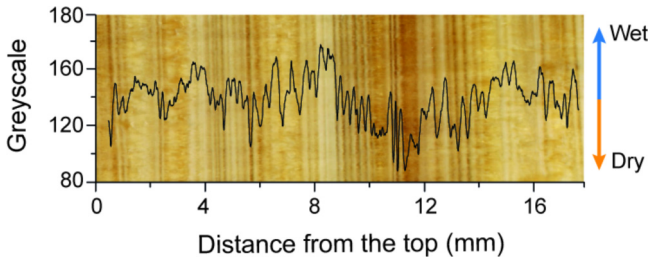


FIG. 5. Grayscale record (in black) extracted from a high-resolution scan of the surface of the stalagmite C132 from Niue island. Lower gray values are associated with dense microcrystalline calcite layers that form during drier periods.

arise from crystallographic variations caused by changes in the stalagmite growth rate (Fig. 5). Grayscale values are obtained from high-resolution scans of the stalagmite surface along its growth axis subsequently extracted with ImageJ [61]. During the dry season, low drip rates promote the deposition of layers with compact and dark crystals, yielding low grayscale values. In the wet season, the drip rates are higher, and crystal growth is enhanced as dissolved inorganic carbon is supplied to a greater extent (see Fig. 5). The inferred link between dark layers and dry season is supported by earlier studies [62,63].

Prior to the recurrence analysis of the grayscale time series, we subtracted a centennial-scale trend using a Gaussian kernel filter in order to focus on the high-frequency variability in the record [Fig. 6(a), black line]. Next, we downsampled the time series uniformly by only storing every tenth value due to computational constraints. This downsampling does not alter the relative changes in the sampling rate [Fig. 6(b)]. The number of samples per year (i.e., the segment size) undergoes an abrupt shift at ≈ 6.15 ka BP. The period with the highest average segment size (≈ 6.4 to 6.15 ka BP) coincides with the wettest period covered by the record, indicated by high grayscale values. This suggests that during this wet period, stalagmite growth was enhanced which resulted in thicker crystal laminae and a higher number of samples per layer. This observation reflects the complex nature of irregular sampling of palaeoclimate-proxy data. If spatial sampling on the stalagmite is performed such that the number of samples is as high as possible, it will inevitably be linked to its growth rate and thus to other environmental parameters and their nonstationary characteristics. Finally, we perform the recurrence analysis [Fig. 6(c)]. In order to characterize seasonal features, the period covered by one segment is fixed as one year. Optimization of deletion and adding costs yields $\Lambda_s = 2$. A window size of $s = 200$ years is chosen with 90% overlap. A recurrence plot with fixed recurrence rate of 15% is obtained for each window and analyzed with DET. DET reveals variations in seasonal-scale predictability for the real grayscale record [Fig. 6(c), blue line]. The effect of the varying sampling size is obtained by the 95% quantile of the DET distribution from 200 SRC surrogates [Fig. 6(c), yellow line]. Five exemplar SRC-surrogate realizations are shown in Appendix B. Both DET time series indicate an increase of seasonal-scale predictability during the wet period between 6.35 and 6.2 ka BP, potentially caused by the simultaneously

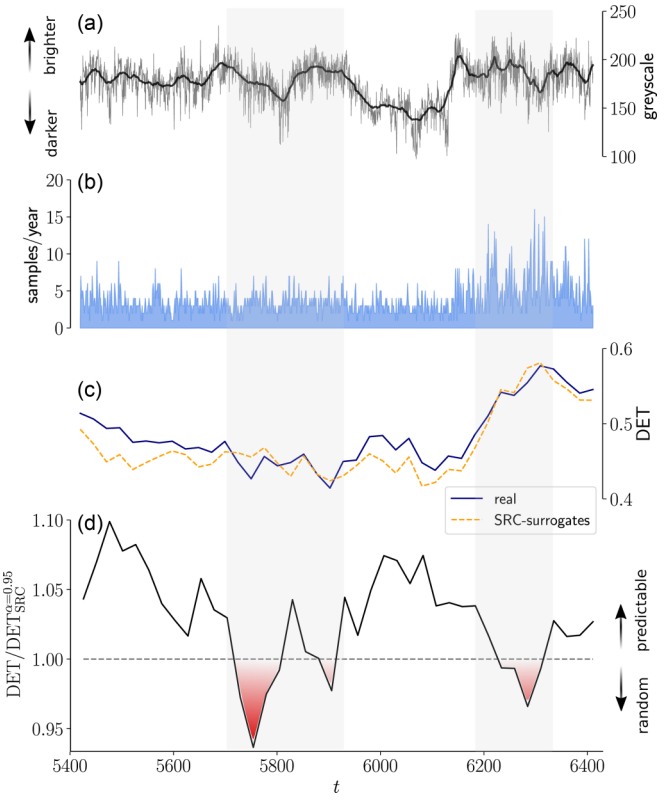


FIG. 6. Application of SRC-surrogate correction method to an irregularly sampled grayscale proxy record from the central Pacific (a) with nonstationary sampling rate (blue) (b). A sliding window RQA using determinism (DET) as a predictability measure is carried out (c). Real DET values are displayed in dark blue, the 95% confidence level computed from 200 SRC surrogates is shown in yellow. Dividing $\mathrm{DET}_{\mathrm{real}}$ by $\mathrm{DET}_{\mathrm{surr}}$ yields a sampling-bias corrected predictability series (d) that indicate mid-Holocene variations in seasonal-scale predictability. Gray shading indicates two periods with low seasonal predictability.

increased sampling rate. The predictability estimate is corrected for the identified sampling bias by considering the ratio $\mathrm{DET_r}/\mathrm{DET_{surr}}$ [Fig. 6(d)]. Two periods (6.4 and 6.2 ka BP, and between 5.9 and 5.72 ka BP) show relatively low segment size-corrected seasonal predictability $\mathrm{DET_r}/\mathrm{DET_{surr}}<1$. While the latter is not significantly affected by the correction, the former can only be identified as less predictable when the variations in sampling rate are taken into account. This result corroborates previous findings that suggested that both of these identified periods were more irregular, i.e., showing less steady seasonal fluctuations [63]. However, it was not possible to characterize all subannual values as a proxy for subannual rainfall distribution rather extracting only the contrast between wet and dry season. The (m)Edit-distance approach employed here in combination with the proposed correction technique allows for a more reliable interpretation of mid-Holocene seasonal variations in the west Pacific.

In particular, an enhanced control of the seasonal cycle by variability was found in [63] for the periods of reduced predictability (6.4 and 6.2 ka BP, and between 5.9 and 5.72 ka BP). High tropical cyclone activity between 6.4–6.2 ka BP could have been triggered by increased El Niño-Southern Oscillation (ENSO) activity, yielding a more irregular subannual distribution of rainfall. Our results indicate that not only contrast between both seasons is rendered less predictable during this period but also the seasonal rainfall distribution appears less stable from one year to another. Reconstructing past climate variability at seasonal scale plays a critical role in the context of human adaption to continuous and abrupt climate variations, and therefore our approach has direct relevance for teasing out the seasonal-scale signals.

VI. CONCLUSION

The characterization of time series from complex nonlinear systems is a challenging task. Irregular sampling, i.e., variations in the sampling interval between consecutive values, additionally impedes typical research objectives such as spectral analysis, persistence estimation or quantifying the predictability of a system. Even though interpolation techniques offer a seemingly efficient way of preprocessing a time series to allow application of standard time series analysis tools, these entail various biases which are difficult to control. A different perspective is pursued by the (m)Edit-distance method. Many analysis methods are based on an estimate of (dis)similarity. With the (m)Edit distance, a suitable distance measure between states of a system at different times i and j is defined by computing the transformation cost of segments centered at these time instances. First analyses demonstrated its scope in the context of recurrence analysis, enabling researchers to examine predictability variations of irregularly sampled palaeoclimate time series. Applications to other complex systems (also for time series with "missing values") and other methodological frameworks (e.g., complex networks, clustering, correlation analysis, etc.) are possible and should be attempted in the near future.

For some real-world systems, it is instructive to quantitatively compare sequences corresponding to a specific timescale in order to analyze the scale-specific variability. In such cases, segment sizes will vary in the presence of irregular sampling. Furthermore, splitting time series with a nonstationary sampling rate into segments that do not cover the same time period will result in a mixing of timescales. We have shown that (m)Edit-distance-based recurrence analyses are affected by variations in segment sizes, resulting in a nontrivial sampling bias if episodes with variable sampling rate are included in a single RP. The (m)Edit distance regards pairs of longer segments to be generally more dissimilar than shorter segments due to higher deletion costs. Shifting costs conversely decrease for increasing segment size, resulting in a nontrivial dependence of costs on local sampling rates. When including amplitudes of a signal into the (m)Edit-distance computation, similar general tendencies were observed but the strength of the segment size dependencies varied for different systems. A more detailed examination of how dissimilar amplitude segments of different paradigmatic systems depend on their timescale will be investigated in a future study.

We introduced a numerical technique based on constrained randomization to address and correct the issue of segment-size dependence in recurrence analysis. This method involves generating an ensemble of sampling rate constrained surrogate

realizations (SRC surrogates). Each SRC surrogate reproduces the real variations of sampling rate and its assignment to the corresponding amplitude differences, allowing the ensemble to be used for correcting the undesired segment size dependence. The effectiveness of the proposed correction was applied to a synthetic AR(1)-time series and real palaeo-proxy data. In both applications, a recurrence analysis successfully recovered variations in the scale-specific predictability of the system while discarding spurious effects imprinted by sampling rate variations. We found that seasonal-scale predictability varied significantly during the mid-Holocene in the west Pacific, corroborating and extending the results from a recent study. The reasons for these changes in predictability warrant further investigation.

The identified sampling bias is a specific case of a more general problem in time series analysis; sliding window analyses (or the study of short time series) often suffer from finite-sampling biases which may introduce artificial variability into any statistical indicator that is computed. As pointed out in Sec. V, finite-sampling biases are also not limited to irregular temporal sampling but are likely to also occur in settings where other parameter axes determine suitable window sizes or adaptive windowing is required. In future, the proposed method could also be applied in such settings to test its effectiveness beyond the examples considered in this study. Python code for the generation of SRC surrogates is available at [64].

ACKNOWLEDGMENTS

This research was supported by the Deutsche Forschungsgemeinschaft in the context of the DFG Project No. MA4759/11-1 "Nonlinear empirical mode analysis of complex systems: Development of general approach and application in climate." It also received financial support from the European Union's Horizon 2020 Research and Innovation program (Marie Sklodowska-Curie Grant Agreement No. 691037). D.E. acknowledges funding by TÜBİTAK (Grant No. 118C236) and the BAGEP Award of the Science Academy. C.N.F. acknowledges financial support from the German Academic Exchange Service (DAAD). A.H. acknowledges support from the Royal Society of New Zealand (Grant No. RIS-UOW1501), and the Rutherford Discovery Fellowship program (Grant No. RDF-UOW1601). The authors declare that they have no conflict of interest.

APPENDIX A: ANALYTICAL AND NUMERICAL SEGMENT SIZE RELATIONS

In the following, we elaborate on the systematic dependence of the (m)Edit distance on segment lengths $N_a = |\mathcal{S}_a|, N_b = |\mathcal{S}_b|$ in the most simple application: we study events (i.e., no assumptions about the amplitude of the signal) which are unevenly spaced by exponentially distributed sampling intervals Δ with a sampling rate λ:

$$p(\Delta, \lambda) = \lambda \mathrm{e}^{-\lambda\Delta}. \tag{A1}$$

Consequently, the number of samples per unit interval k is Poisson distributed:

$$\rho(k, \lambda) = \frac{\lambda^k e^{-k}}{k!} \tag{A2}$$

with λ being equivalent to the expected number of samples per unit interval; $\mathbb{E}(X) = \lambda$. Furthermore, the nth time step is Erlang distributed with the rate parameter λ:

$$f(t; n, \lambda) = \frac{\lambda^n t^{n-1} e^{-\lambda t}}{(n-1)!}, \tag{A3}$$

which is a general result for a sum of n independent exponential random variables with equivalent rate parameters λ [65].

We are interested in the segment size dependence of deletion (adding) and shifting costs for the edit distance. This can be evaluated by considering M exponential random variables where each is drawn from a distribution $p(\Delta, \lambda_m)$ with distinct $\lambda_m, m = 1, 2, \ldots, M$. When applied, this setting can be considered equivalent to a scenario where a time axis changes its local sampling rate λ_m at M points and segments from these should be compared via the edit distance. For a specific pair of segments with sizes N_a, N_b, the minimum deletion cost (no deletions as competing to shifts included) for their transformation is

$$C_{\mathrm{del}}(N_a, N_b) = \Lambda_{\mathrm{S}} |N_a - N_b|. \tag{A4}$$

Consequently, for two segments of average sizes $\mathbb{E}[N_a] = \lambda_1, \mathbb{E}[N_b] = \lambda_2$ we obtain a minimum deletion cost of $C_{\mathrm{del}}(\mathbb{E}(N_a), \mathbb{E}(N_b)) = \Lambda_{\mathrm{S}}|\lambda_1 - \lambda_2|$. A cost matrix $C_{\mathrm{del}}(\lambda_1, \lambda_2)$ is exemplified in Fig. 7(a). The expected minimum deletion cost for two randomly selected segments from time periods with different rates λ_1, λ_2 can be computed by using the the Skellam distribution

$$\rho_s(k = |z|; \lambda_1, \lambda_2) = \begin{cases} e^{-\lambda_1-\lambda_2}\left(\left(\frac{\lambda_1}{\lambda_2}\right)^{\frac{k}{2}} I_k(2\sqrt{\lambda_1\lambda_2}) + \left(\frac{\lambda_2}{\lambda_1}\right)^{\frac{k}{2}} I_{-k}(2\sqrt{\lambda_1\lambda_2})\right) & \text{if} k > 0 \\ e^{-\lambda_1-\lambda_2} I_0(2\sqrt{\lambda_1\lambda_2}) & \text{if} k = 0 \end{cases} \tag{A5}$$

for the difference $Z = X - Y$ where X, Y are Poisson-distributed random variables with rates λ_1, λ_2, Eq. (A2). $I_k(a)$ denotes the modified Bessel function of the first kind. For $k > 0$, the moment-generating function is consequently given by

$$M(t; \lambda_1, \lambda_2) = e^{-\lambda_1-\lambda_2}\left(\sum_{k=0}^{\infty} e^{tk} I_k(2\sqrt{\lambda_1\lambda_2}) \times \left[\left(\frac{\lambda_1}{\lambda_2}^{k/2}\right) + \left(\frac{\lambda_2}{\lambda_2}^{k/2}\right)\right] - I_0(2\sqrt{\lambda_1\lambda_2})\right). \tag{A6}$$

With "Marcum's Q"

$$Q(\sqrt{2\lambda_1}, \sqrt{2\lambda_2}) = e^{-\lambda_1-\lambda_2} \sum_{k=0}^{\infty} \left(\frac{\lambda_1}{\lambda_2}\right)^{\frac{k}{2}} I_k(2\sqrt{\lambda_1\lambda_2}) \tag{A7}$$

and its derivative

$$\frac{\mathrm{d}}{\mathrm{d}t} Q(\sqrt{2\lambda_1}, \sqrt{2\lambda_2}) = e^{-\lambda_1 e^t - \lambda_2 e^{-t}}\left[\lambda_2 e^{-t} I_0(2\sqrt{\lambda_1\lambda_2}) + \sqrt{\lambda_1\lambda_2} I_1(2\sqrt{\lambda_1\lambda_2})\right] \tag{A8}$$

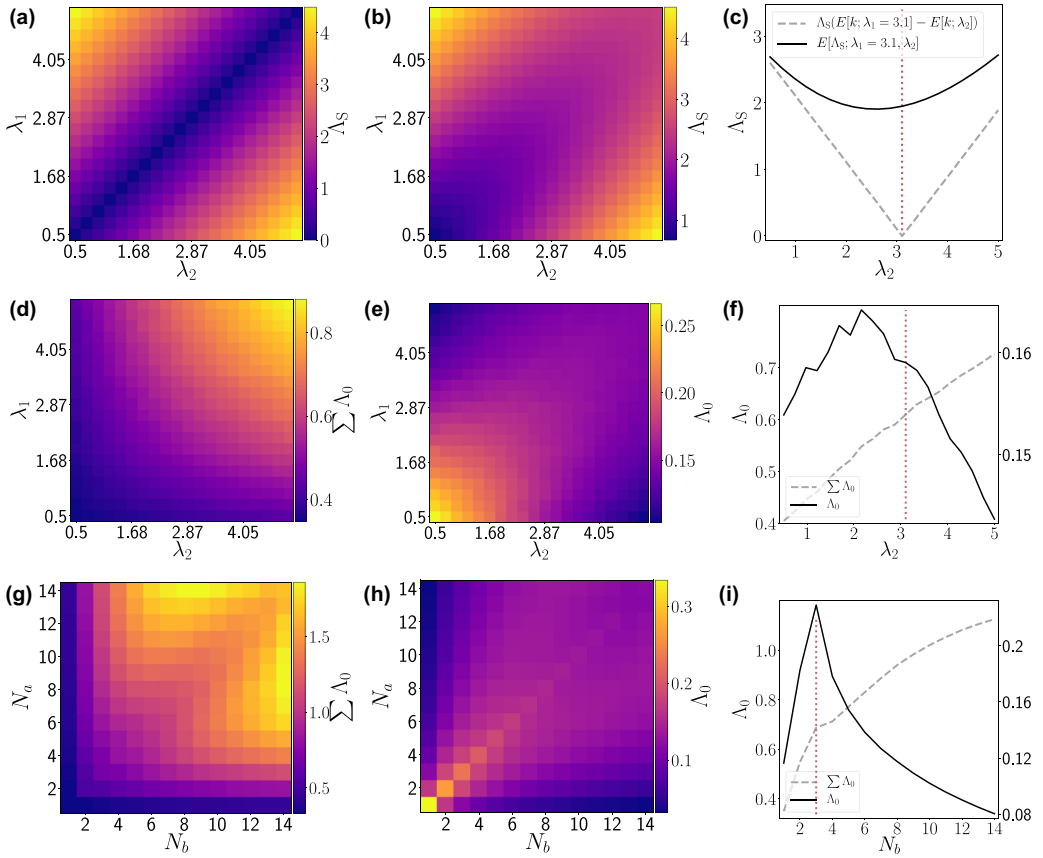


FIG. 7. Subcosts for adding or deleting (a)–(c) and shifting (d)–(i) operations for exponentially distributed sampling intervals. Only necessary deleting or adding operations are regarded ("no competing operations"). The sampling rate dependence of deletion costs is given as difference between expected number of samples per unit interval (left matrix and gray dashed line) and as expected costs given two rates λ_1, λ_2 [right matrix and black line, Eq. (A10)]. Shifting costs are studied (b) numerically with respect to their dependence on the sampling rates λ_1, λ_2 and (c) on the actual number of samples per unit interval N_a, N_b. The left matrices shows shifting costs for the full transformation of segments, the center matrices show shifting costs per operation. Exemplary columns are displayed in the line plots, whereas the red dashed line marks the respective rate λ_1 and segment size N_b.

this can be written as

$$M(t;\lambda_1,\lambda_2) = e^{-\lambda_1-\lambda_2}\Big[Q\Big(\sqrt{2\lambda_2 e^{-t}}, \sqrt{2\lambda_1 e^{t}}\Big)e^{\lambda_1 e^t+\lambda_2 e^{-t}} \times Q\Big(\sqrt{2\lambda_1 e^{-t}}, \sqrt{2\lambda_2 e^{t}}\Big)e^{\lambda_2 e^t+\lambda_1 e^{-t}} I_0(2\sqrt{\lambda_1\lambda_2})\Big]. \tag{A9}$$

Differentiating this moment-generating function [using Eq. (A8)] around $t = 0$ with Leibniz rule yields the expected value:

$$\mathbb{E}[k;\lambda_1,\lambda_2] = 2e^{-\lambda_1-\lambda_2}\Big[\lambda_2 I_0(2\sqrt{\lambda_1\lambda_2}) + \sqrt{\lambda_1\lambda_2} I_1(2\sqrt{\lambda_1\lambda_2})\Big] + [(\lambda_2-\lambda_1)(1-2Q(\sqrt{2\lambda_1}, \sqrt{2\lambda_2}))]. \tag{A10}$$

Hence, $\mathbb{E}[C_{\text{del}}(\lambda_1,\lambda_2)] = \Lambda_S \mathbb{E}[k;\lambda_1,\lambda_2]$ [Fig. 7(a), middle]. In the right line plot of Fig. 7(a), two columns with λ_1 fixed at 3.1 are shown to illustrate the scaling of deletion costs with the rate λ more clearly. While $C_{\text{del}}(\mathbb{E}(N_a), \mathbb{E}(N_b))$ shows a sharp minimum at the rate $\lambda_2 = \lambda_1$, $\mathbb{E}[k;\lambda_1,\lambda_2]$ decreases more smoothly with increasing λ_2, and increases afterwards. The latter becomes minimal for a value $\lambda < \lambda_2$ instead of $\lambda = \lambda_2$ since Poisson distributions $\rho(k,\lambda)$ are right-skewed, having higher cumulated probability mass for all values $k > \lambda$. Note that all said above holds in the same way for adding operations.

For the analysis of shifting costs, we focus on the simple case of linear shifting costs

$$\tilde{f}_{\Lambda_0}(t(\alpha), t(\beta)) = |t(\alpha) - t(\beta)| \tag{A11}$$

between the αth event in segment $\mathcal{S}_a$ and the βth event in segment $\mathcal{S}_b$ as proposed in the original, unmodified edit-distance measure. To exclude effects caused by absolute timing of events, timing of events within each segment is always transformed into the interval $I = [0, 1]$. The sum of all shifting costs for a pair of segments is denoted as $d_{ab} = \Lambda_0 \sum_{\alpha,\beta} f_{\Lambda_0}(t(\alpha), t(\beta))$ with $\Lambda_0 = 1$. Note that $N_a = N_b$ as $|N_a - N_b|$ deletions and additions have already been carried out. A closed-form solution for the shifting costs between two time instances drawn randomly from the distributions $f(x;m_1,\lambda_1)$, $f(y;m_2,\lambda_2)$ most likely exists, at least for the case $m_1 = m_2$ but its computation is beyond this study. We examine shifting costs for this case numerically, while we explicitly exclude any deletions as an alternative operation

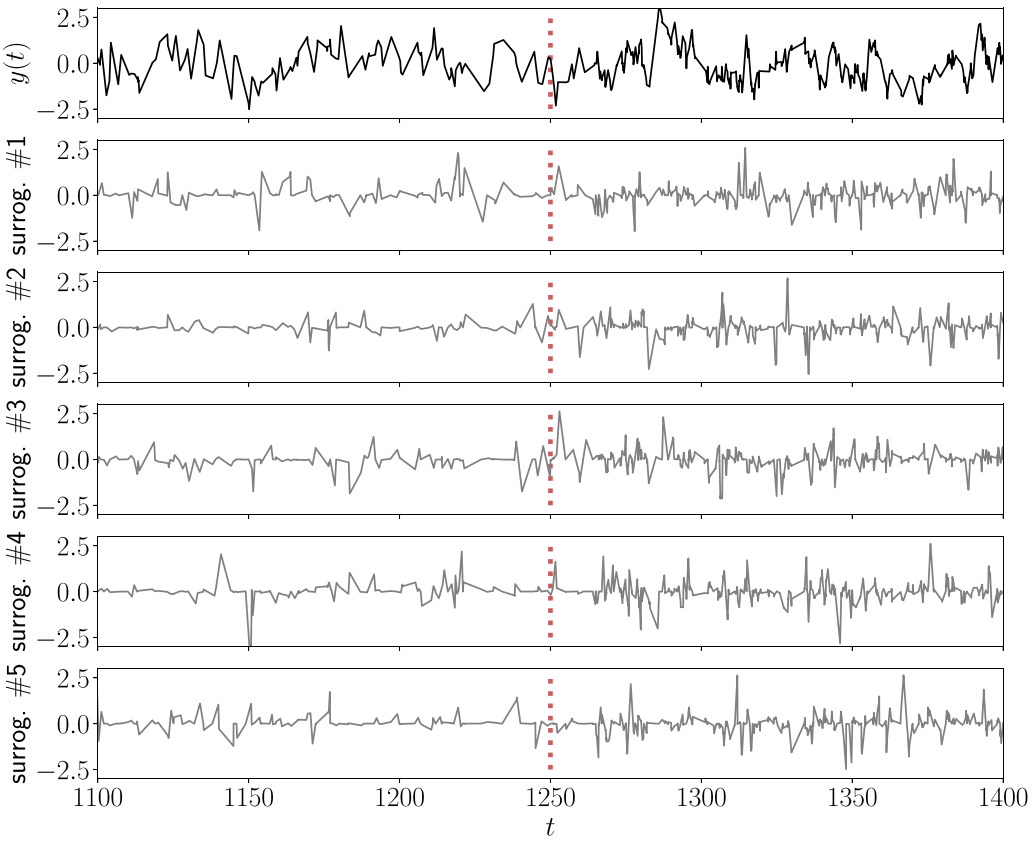


FIG. 8. Zoomed section of synthetic AR(1)-time series (black) and five exemplar SRC-surrogate realizations (gray). The red dotted line indicates the transition of the sampling rate towards more dense sampling. Sampling intervals are γ-distributed.

to shifting after the necessary $|N_a - N_b|$ deletions ("basic deletions") [Fig. 7(b)]. We fix $w = 1$ as the unit interval (arbitrary units). The numerical estimate of the average cost for transforming a segment sampled with rate λ_1 into a segment sampled with rate λ_2 is based on generating time axes for a fixed time period $T = 10\,000$ (but varying number of events). Given a fixed combination of λ_1, λ_2, a total of $K = 10\,000$ segment pairs are randomly sampled (with replacement) from both corresponding time axes. The edit distance is computed for each pair of segments and averaged over all pairs to obtain a single value $\overline{d}(\lambda_1, \lambda_2)$ that is characteristic for the combination of rates λ_1, λ_2. This is shown as a cost matrix $C_{\text{shift}}(\lambda_1, \lambda_2)$ of averaged total shifting costs between randomly drawn segments [Fig. 7(b), left]. The total number of shifts performed after deleting $|N_a - N_b|$ events generally differs for distinct pairs of segments $\mathcal{S}_a, \mathcal{S}_b$ at fixed λ_1, λ_2. However, when averaged over all randomly drawn segment pairs, an increasing trend along the diagonals is observed. Furthermore, average total shifting costs $\overline{d}(\lambda_1, \lambda_2)$ increase for fixed λ_1 and increasing λ_2 [Fig. 7(b), right] which is to be expected as a higher number of shifts will entail higher summed costs. On the other hand, no monotonous relation between the average shifting costs per shifting operation

$$\tilde{C}_{\text{shift}}(\lambda_1, \lambda_2) = \sum_{k=1}^{K} d\left(\mathcal{S}_{a,k}^{(\lambda_1)}, \mathcal{S}_{b,k}^{(\lambda_2)}\right) \Big/ \max\{N_a, N_b\}, \quad \text{(A12)}$$

and sampling rate is observed [Fig. 7(b), center and bold black line on the right]. With increasing sampling rates, the cost of an average single shifting operation decreases (diagonals of the matrix). For fixed λ_1, it is maximized at a value $\lambda_2 < \lambda_1$ for the same reason as above, i.e., the Erlang distribution being right-skewed.

If we instead examine the dependence of shifting costs on the actual segment size $\tilde{C}_{\text{shift}}(N_a, N_b)$ rather than the rates [Fig. 7(c)], a sharp maximum at $N_b = \lambda_1$ is found (black line, right plot). Total shifting costs increase for $N_b < \lambda_1$ and continue to increase more slowly for $N_b > \lambda_1$. For fixed N_a, an increasing number N_b of events per unit interval increases the likelihood that some events are placed close to the events in segment $\mathcal{S}_a$, resulting in lower distances $d_{ab}(N_a, N_b)$.

APPENDIX B: SAMPLING RATE-CONSTRAINED SURROGATES

The proposed sampling-rate correction approach involves a constrained randomization procedure, in which sampling rate-constrained surrogates (SRC surrogates) are generated. To illustrate the resulting time series, we show five SRC-surrogate realizations of the irregularly sampled AR(1)-process from Sec. IV B in Fig. 8. The transition in sampling rate (dotted red line) is well visible from the different surrogate realizations.

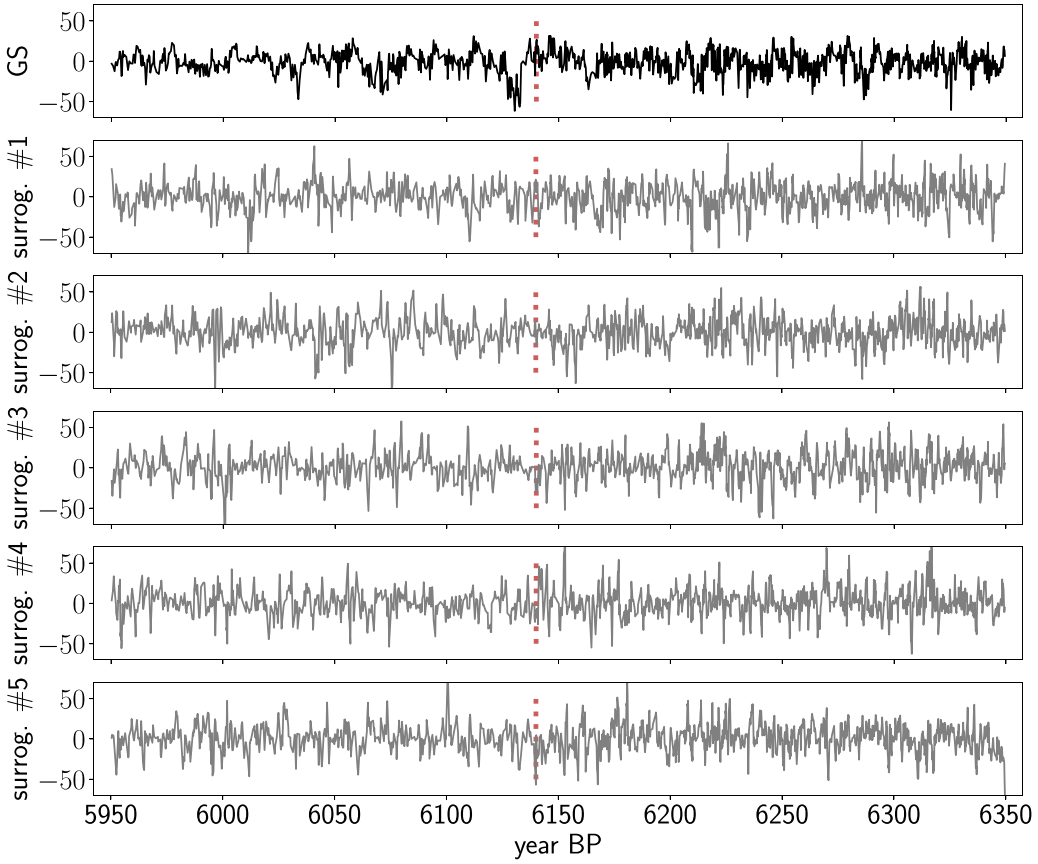


FIG. 9. Zoomed section of grayscale anomaly time series (black) and five exemplar SRC-surrogate realizations (gray). The red dotted line indicates the transition of the sampling rate towards more dense sampling.

We can also identify the rapid increase in sampling rate for the grayscale proxy time series in the real-world example from Sec. V (Fig. 9). Visually, it is expressed as an increase of variance which is reproduced by the SRC-surrogate realizations.

[1] R. Renò, A closer look at the Epps effect, Int. J. Theor. Appl. Finance **6**, 87 (2003).

[2] A. Fedotov, S. Akulov, and E. Timchenko, Methods of mathematical analysis of heart rate variability, Biomed. Eng. **54**, 220 (2020).

[3] C. K. Enders, *Applied Missing Data Analysis* (Guilford Press, 2010).

[4] M. Khayati, A. Lerner, Z. Tymchenko, and P. Cudré-Mauroux, Mind the gap: An experimental evaluation of imputation of missing values techniques in time series, Proc. VLDB Endow. **13**, 768 (2020).

[5] J. D. Scargle, Studies in astronomical time series analysis. II—Statistical aspects of spectral analysis of unevenly spaced data, Astrophys. J. **263**, 835 (1982).

[6] K. Rehfeld, N. Marwan, J. Heitzig, and J. Kurths, Comparison of correlation analysis techniques for irregularly sampled time series, Nonlin. Process. Geophys. **18**, 389 (2011).

[7] J. U. Baldini, Detecting and quantifying paleoseasonality in stalagmites using geochemical and modelling approaches, AGUFM **2017**, PP54B–09 (2021).

[8] C. Mühlinghaus, D. Scholz, and A. Mangini, Modelling stalagmite growth and $\delta^{13}C$ as a function of drip interval and temperature, Geochim. Cosmochim. Acta **71**, 2780 (2007).

[9] J. Garland, T. R. Jones, M. Neuder, V. Morris, J. W. White, and E. Bradley, Anomaly detection in paleoclimate records using permutation entropy, Entropy **20**, 931 (2018).

[10] M. H. Trauth, Spectral analysis in quaternary sciences, Quaternary Sci. Rev. **270**, 107157 (2021).

[11] K. Rehfeld, N. Marwan, S. F. Breitenbach, and J. Kurths, Late Holocene Asian summer monsoon dynamics from small but complex networks of paleoclimate data, Climate Dyn. **41**, 3 (2013).

[12] M. Schulz and K. Stattegger, Spectrum: Spectral analysis of unevenly spaced paleoclimatic time series, Comput. Geosci. **23**, 929 (1997).

[13] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, Smote: Synthetic minority over-sampling technique, J. Artif. Intell. Res. **16**, 321 (2002).

[14] S. Barua, M. M. Islam, X. Yao, and K. Murase, MWMOTE—Majority weighted minority oversampling technique for

imbalanced data set learning, IEEE Trans. Knowl. Data Eng. **26**, 405 (2012).

[15] T. M. Lenton, J. Rockström, O. Gaffney, S. Rahmstorf, K. Richardson, W. Steffen, and H. J. Schellnhuber, Climate tipping points—Too risky to bet against, Nature (London) **575**, (2019).

[16] E. Bradley and H. Kantz, Nonlinear time-series analysis revisited, Chaos **25**, 097610 (2015).

[17] N. Marwan, J. F. Donges, R. V. Donner, and D. Eroglu, Nonlinear time series analysis of palaeoclimate proxy records, Quat. Sci. Rev. **274**, 107245 (2021).

[18] M. Scheffer, J. Bascompte, W. A. Brock, V. Brovkin, S. R. Carpenter, V. Dakos, H. Held, E. H. Van Nes, M. Rietkerk, and G. Sugihara, Early-warning signals for critical transitions, Nature (London) **461**, 53 (2009).

[19] J. Lekscha and R. V. Donner, Phase space reconstruction for non-uniformly sampled noisy time series, Chaos **28**, 085702 (2018).

[20] M. McCullough, K. Sakellariou, T. Stemler, and M. Small, Counting forbidden patterns in irregularly sampled time series. I. The effects of under-sampling, random depletion, and timing jitter, Chaos **26**, 123103 (2016).

[21] K. Sakellariou, M. McCullough, T. Stemler, and M. Small, Counting forbidden patterns in irregularly sampled time series. II. Reliability in the presence of highly irregular sampling, Chaos **26**, 123104 (2016).

[22] S. Suzuki, Y. Hirata, and K. Aihara, Definition of distance for marked point process data and its application to recurrence plot-based analysis of exchange tick data of foreign currencies, Int. J. Bifurcation Chaos **20**, 3699 (2010).

[23] I. 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).

[24] E. Ukkonen, Algorithms for approximate string matching, Inf. Control **64**, 100 (1985).

[25] A. Banerjee, B. Goswami, Y. Hirata, D. Eroglu, B. Merz, J. Kurths, and N. Marwan, Recurrence analysis of extreme event like data, in *Nonlinear Processes in Geophysics Discussions*, edited by T. Miyoshi (Copernicus Publications for the European Geosciences Union, Göttingen, 2020), pp. 1–25.

[26] N. Marwan, M. C. Romano, M. Thiel, and J. Kurths, Recurrence plots for the analysis of complex systems, Phys. Rep. **438**, 237 (2007).

[27] E. Garcia-Ceja, M. Z. Uddin, and J. Torresen, Classification of recurrence plots' distance matrices with a convolutional neural network for activity recognition, Proc. Comput. Sci. **130**, 157 (2018).

[28] M. C. Romano, M. Thiel, J. Kurths, and W. von Bloh, Multivariate recurrence plots, Phys. Lett. A **330**, 214 (2004).

[29] N. Marwan, S. Schinkel, and J. Kurths, Recurrence plots 25 years later—Gaining confidence in dynamical transitions, Europhys. Lett. **101**, 20007 (2013).

[30] G. Corso, T. d. L. Prado, G. Z. d. S. Lima, J. Kurths, and S. R. Lopes, Quantifying entropy using recurrence matrix microstates, Chaos **28**, 083108 (2018).

[31] S. Schinkel, N. Marwan, and J. Kurths, Order patterns recurrence plots in the analysis of ERP data, Cogn. Neurodyn. **1**, 317 (2007).

[32] T. Braun, V. R. Unni, R. I. Sujith, J. Kurths, and N. Marwan, Detection of dynamical regime transitions with lacunarity as a multiscale recurrence quantification measure, Nonlin. Dyn. **104**, 3955 (2021).

[33] I. 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).

[34] N. Marwan, D. Eroglu, I. Ozken, T. Stemler, K.-H. Wyrwoll, and J. Kurths, Regime change detection in irregularly sampled time series, in *Advances in Nonlinear Geosciences* (Springer International Publishing, Basel, 2018), pp. 357–368.

[35] M. A. Stegner, Z. Ratajczak, S. R. Carpenter, and J. W. Williams, Inferring critical transitions in paleoecological time series with irregular sampling and variable time-averaging, Quat. Sci. Rev. **207**, 49 (2019).

[36] K.-H. Krämer, G. Datseris, 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).

[37] K. Rehfeld and J. Kurths, Similarity estimators for irregular and age-uncertain time series, Climate Past **10**, 107 (2014).

[38] M. Mudelsee, *Climate Time Series Analysis* (Springer, Berlin, 2013).

[39] W. J. Masek and M. S. Paterson, A faster algorithm computing string edit distances, J. Comput. Syst. Sci. **20**, 18 (1980).

[40] L. Rabiner, A. Rosenberg, and S. Levinson, Considerations in dynamic time warping algorithms for discrete word recognition, IEEE Trans. Acoust. Speech, Signal Process. **26**, 575 (1978).

[41] J. D. Victor and K. P. Purpura, Metric-space analysis of spike trains: Theory, algorithms and application, Netw. Comput. Neural Syst. **8**, 127 (1997).

[42] J. O. Eckmann, S. Oliffson Kamphorts, and D. Ruelle, Recurrence plots of dynamical systems, Europhys. Lett. **4**, 973 (1987).

[43] H. Poincaré, Sur le problème des trois corps et les équations de la dynamique, Acta Math. **13**, A3 (1890).

[44] K. H. Kraemer, R. V. Donner, J. Heitzig, and N. Marwan, Recurrence threshold selection for obtaining robust recurrence characteristics in different embedding dimensions, Chaos **28**, 085720 (2018).

[45] S. Schinkel, O. Dimigen, and N. Marwan, Selection of recurrence threshold for signal detection, Eur. Phys. J.: Spec. Top. **164**, 45 (2008).

[46] Y. Hirata, Recurrence plots for characterizing random dynamical systems, Commun. Nonlin. Sci. Num. Simul. **94**, 105552 (2021).

[47] A. M. T. Ramos, A. Builes-Jaramillo, G. Poveda, B. Goswami, E. E. N. Macau, J. Kurths, and N. Marwan, Recurrence measure of conditional dependence and applications, Phys. Rev. E **95**, 052206 (2017).

[48] T. Westerhold, N. Marwan, A. J. Drury, D. Liebrand, C. Agnini, E. Anagnostou, J. S. Barnet, S. M. Bohaty, D. De Vleeschouwer, F. Florindo *et al.*, An astronomically dated record of Earth's climate and its predictability over the last 66 million years, Science **369**, 1383 (2020).

[49] H. Kantz, A robust method to estimate the maximal Lyapunov exponent of a time series, Phys. Lett. A **185**, 77 (1994).

[50] *Nonlinear Time Series Analysis in the Geosciences: Applications in Climatology, Geodynamics and Solar-Terrestrial Physics*, edited by R. V. Donner and S. M. Barbosa, Lecture Notes in Earth Sciences Vol. 112 (Springer, Berlin, 2008).

[51] D. 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).
[52] T. March, S. Chapman, and R. Dendy, Recurrence plot statistics and the effect of embedding, Physica D **200**, 171 (2005).
[53] D. C. Williams, Finite sample correction factors for several simple robust estimators of normal standard deviation, J. Stat. Comput. Simul. **81**, 1697 (2011).
[54] C. Park, H. Kim, and M. Wang, Investigation of finite-sample properties of robust location and scale estimators, Communications in Statistics-Simulation and Computation (2019), pp. 1–27.
[55] K. E. Trenberth, Some effects of finite sample size and persistence on meteorological statistics. Part I: Autocorrelations, Month. Weather Rev. **112**, 2359 (1984).
[56] B. Goswami, P. Schultz, B. Heinze, N. Marwan, B. Bodirsky, H. Lotze-Campen, and J. Kurths, Inferring interdependencies from short time series, Ind. Acad. Sci. Conf. Ser. **1**, 51 (2017).
[57] T. Schreiber and A. Schmitz, Surrogate time series, Physica D **142**, 346 (2000).
[58] X. Luo, T. Nakamura, and M. Small, Surrogate test to distinguish between chaotic and pseudoperiodic time series, Phys. Rev. E **71**, 026230 (2005).
[59] G. Lancaster, D. Iatsenko, A. Pidde, V. Ticcinelli, and A. Stefanovska, Surrogate data for hypothesis testing of physical systems, Phys. Rep. **748**, 1 (2018).
[60] T. Schreiber, Constrained randomization of time series data, Phys. Rev. Lett. **80**, 2105 (1998).
[61] M. D. Abràmoff, P. J. Magalhães, and S. J. Ram, Image processing with ImageJ, Biophotonics In. **11**, 36 (2004).
[62] P. Aharon, M. Rasbury, and V. Murgulet, Caves of Niue Island, South Pacific: Speleothems and water geochemistry, Geol. Soc. Amer. Special Papers **404**, 283 (2006).
[63] C. Nava-Fernandez, T. Braun, B. Fox, A. Hartland, O. Kwiecien, C. Pederson, S. Höpker, S. Bernasconi, M. Jaggi, J. Hellstrom *et al.*, Mid-Holocene rainfall changes in the southwestern Pacific, Climate of the Past Discussions (2022), pp. 1–29.
[64] https://github.com/ToBraun/SRC-surrogates.
[65] D. R. Cox, *Renewal Theory* (Springer, Berlin, 1967).