Journal of Physics: Complexity

OPEN ACCESS

PERSPECTIVE

Complex systems approaches for Earth system data analysis

Niklas Boers[1,2,3,7,*], Jürgen Kurths[1,4,5,7] and Norbert Marwan[1,6,7]

[1] Potsdam Institute for Climate Impact Research, Potsdam, Germany
[2] Department of Mathematics and Computer Science, Freie Universität Berlin, Berlin, Germany
[3] Department of Mathematics and Global Systems Institute, University of Exeter, Exeter, United Kingdom
[4] Department of Physics, Humboldt University Berlin, Berlin, Germany
[5] Nizhny Novgorod State University, Nizhny Novgorod, Russia
[6] Institute of Geosciences, University of Potsdam, Potsdam, Germany
[7] These authors contributed equally to this work.
[*] Author to whom any correspondence should be addressed.

E-mail: boers@pik-potsdam.de

Keywords: complexity science, data analysis, complex networks, recurrence

RECEIVED
30 October 2020

REVISED
21 December 2020

ACCEPTED FOR PUBLICATION
6 January 2021

PUBLISHED
8 April 2021

Abstract

Complex systems can, to a first approximation, be characterized by the fact that their dynamics emerging at the macroscopic level cannot be easily explained from the microscopic dynamics of the individual constituents of the system. This property of complex systems can be identified in virtually all natural systems surrounding us, but also in many social, economic, and technological systems. The defining characteristics of complex systems imply that their dynamics can often only be captured from the analysis of simulated or observed data. Here, we summarize recent advances in nonlinear data analysis of both simulated and real-world complex systems, with a focus on recurrence analysis for the investigation of individual or small sets of time series, and complex networks for the analysis of possibly very large, spatiotemporal datasets. We review and explain the recent success of these two key concepts of complexity science with an emphasis on applications for the analysis of geoscientific and in particular (palaeo-) climate data. In particular, we present several prominent examples where challenging problems in Earth system and climate science have been successfully addressed using recurrence analysis and complex networks. We outline several open questions for future lines of research in the direction of data-based complex system analysis, again with a focus on applications in the Earth sciences, and suggest possible combinations with suitable machine learning approaches. Beyond Earth system analysis, these methods have proven valuable also in many other scientific disciplines, such as neuroscience, physiology, epidemics, or engineering.

1. Introduction

Data analysis is among the oldest techniques used in science and even long before, for various practical purposes. For example, some of the earliest instances are related to the determination of rather regular weather conditions for planning relocations during the hunter-gatherer epochs, or later for planning of sowing and harvesting of agriculture-based sedentary societies. Another important challenge has been the prediction of rare but potentially recurring events, such as Sun eclipses, appearances of a comets, epidemic outbreaks, heat waves, floodings, or storm. Some of these basic problems were solved long ago, while others are still challenging problems today, as we will explain in the following.

One of the properties of natural systems that has been used to address the above mentioned problems is recurrence. Recurrence is a typical feature of dynamical systems that is often observed in our daily life and across all scientific disciplines. Recurring phenomena were well known, attracted also much attention, and were used already in ancient times. Examples are early astronomical observatories (one of the earliest was probably Stonehenge [1, 2]), early explanations of the motions of the planets by Greek philosophers [3], or the Maya calendar, which consists of several calendars of different periods and for different use, but all using a daily

counting on a base of 20 days. A seminal finding by the French mathematician Poincaré finally emphasized the importance of recurrences for the study of dynamical systems and provided a mathematically sound concept: in the context of studying the three-body problem for explaining celestial dynamics, he formulated the recurrence theorem in 1890 which states that a conservative (i.e., volume preserving) dynamical system with bounded orbits returns infinitely many times as close as one wishes to its initial state [4]. A measure-theoretic proof of the theorem was given by Carathéodory in 1919 [5].

A basic consequence of recurrence is the existence of periodicities in natural processes. Some of them, as the one related to the daily or annual cycles, are very distinct. But others are not so clear, as for example those appearing in the sunspot series or in the Canadian hare-lynx data [6]. The latter are often called hidden periodicities. In analogy to optics, Schuster invented in 1898 the periodogram-technique [7] and he argued that '…the periodogram furnishes more definite information than the optical instrument can' [8]. This was the beginning of modern time series analysis. He applied the periodogram to the analysis of sunspot data in 1905 and inferred a mean period of 11.125 yr [8]. Later on, a complete theory of linear statistical time series analysis has been developed, where the description via autoregressive processes by Yule in 1927 was a milestone [9]. There are various textbooks on power spectrum analysis, linear models and filtering, such as Box and Jenkins [10] or Priestley [11] and the corresponding techniques are available in many computational toolboxes.

Although such kind of linear methods for the analysis of time series have found a lot of very successful applications in science, engineering, (socio-)economics, physiology, psychology etc, in the course of the last decades of the 20th century, the limits to linearity have become more and more important. The strong progress in the study of nonlinear dynamical systems in the 1980's and 1990's opened new doors for a more appropriate analysis of complex nonlinear systems, such as lasers, the human brain, power grids, or the Earth system and its components. Techniques for the estimation of basic characteristics of nonlinear systems, such as fractal dimensions, Lyapunov exponents or Kolmogorov entropy, were worked out and applied to various disciplines with great success (cf Kantz and Schreiber [12]). Recurrence properties have been used for this purpose as well and led to recurrence plots (RPs) [13] and recurrence quantification analysis (RQA) [14]. This specific approach is attracting increasing attention and finding applications in many different disciplines [15–17]. For example, it helps in detecting Parkinson's disease from handwriting [18], provided new insights into the impact of palaeoclimate variability on human evolution [19], or uncovered the mechanism in combustion processes leading to unstable and critical states in gas turbines [20]. We present this approach in detail in section 2.

A challenge in investigating complex systems is their intrinsic composition of nonlinearly interlinked subcomponents, requiring the analysis of multivariate nonlinear time series. The above mentioned techniques are mostly also suitable for multivariate time series. However, due to more refined measurement technologies, ever larger data sets are becoming available at an ongoing basis. These datasets, such as the ones from remote-sensing measurements, depend on both time and space, called spatio-temporal data, and require more suitable techniques. One promising way to treat such kind of big data is based on complex networks or graphs. The main reason for the very rapid evolution of complex network science during the last two decades is that it allows a much better description of various real-world processes, as demonstrated first in sociology and engineering, and later also in neuroscience, Earth sciences, and several other fields [21–28].

There is a basic problem to model a continuous dynamical system, as the brain or the climate, with a discrete-in-space structure such as complex networks. As first proposed in 2004 [29], one option is to consider different spatial regions as nodes and then define network links between the different regions based on suitable similarity measures from nonlinear data analysis. In section 3 we will explain this approach and present recent progress to show the great potential of this methodology. In particular, these developments have recently enabled to detect new mechanisms in the coupled eco-climate system in Amazonia and even led to much better predictions of the Indian summer monsoon and El Niño activities.

Finally, we give open challenging problems in both chapters.

2. Recurrence analysis

2.1. State-of-the-art

For the numerical study of recurring processes, several approaches are of interest. The power spectrum analysis is probably one of the best known and widely used techniques for the analysis of periodicities in time series [7], revealing the main periods within the measured signal (figure 1(B)). Wavelet analysis reveals similar information, additionally providing a potential change of the detected periods over time (figure 1(C)). Both approaches are useful, although coming with some limitations, especially in the presence of harmonics, non-

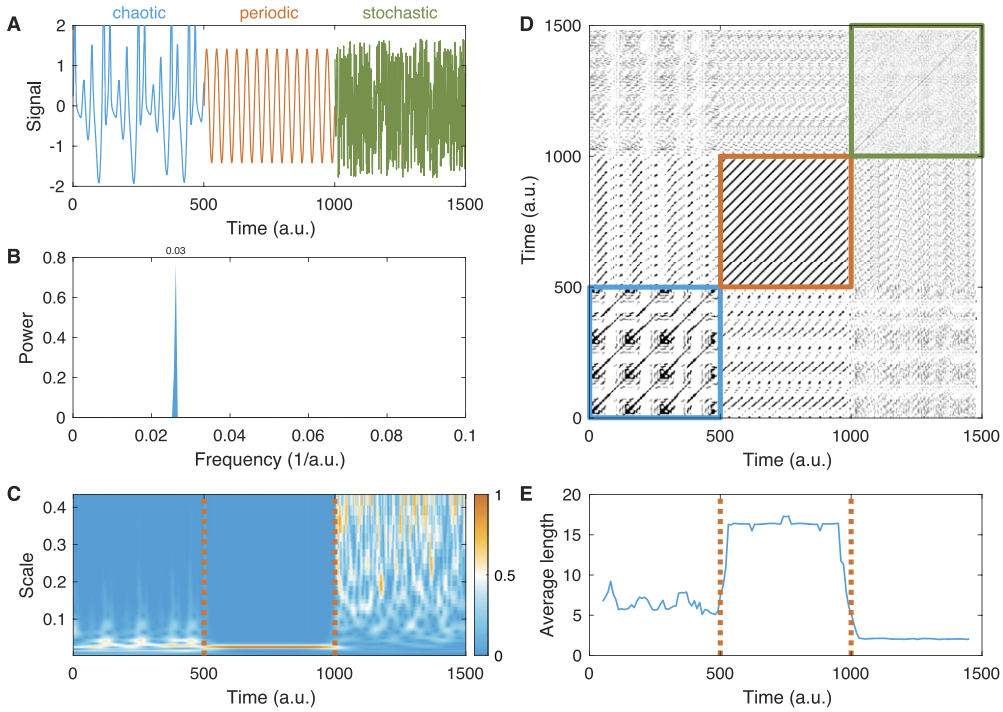

Figure 1. Exemplary RP (D) showing the recurrence structure of a sequence of chaotic, periodic, and stochastic dynamics (A). The power spectrum (B) only reveals the dominating frequency of the periodic epoch, although the signal consists of further frequencies; the wavelet diagram (C) represents the different scales within the three segments of different dynamics. Quantifying the RP (E) allows further insights into the dynamics, such as predictable time as derived from the average line length in the RP.

stationarity, nonlinearity, trends and noise, non-periodic signals, or generalizations for analyzing multi-variate or spatial and spatio-temporal data.

A fundamental approach that can be used to investigate recurring features in time series (and even in spatial data) is the *RP* [13] and its quantification *RQA* [14, 30]. This approach is not restricted to periodic variations and has its roots in the theory of dynamical systems.

A RP is a two-dimensional, discrete, and finite representation of a dynamical system of arbitrary dimension with state vectors $\vec{x}_i$ ($t = i\Delta t$ and sampling time Δt). If the distance $d_{i,j}$ between two states $\vec{x}_i$ and $\vec{x}_j$ at times i and j is very small, then the state $\vec{x}_i$ recurred at time j. $d_{i,j}$ can be defined in different ways, depending on the current research question. A common choice is simply the Euclidean distance between the states $\vec{x}_i$ and $\vec{x}_j$, i.e., $d_{i,j} = \|\vec{x}_i - \vec{x}_j\|$. All pairwise tests for such pairs of recurring states form the square matrix $\mathbf{R}$ (figure 2),

$$R_{i,j} = \Theta(\varepsilon - d_{i,j}), \tag{1}$$

where the recurrence threshold ε finally defines the recurrences. Its graphical representation is called RP and already provides a visual impression of the dynamics of the system (figure 1(D)). It even allows us to look at high-dimensional systems which are almost impossible to visualize by this two-dimensional representation.

If only a univariate time series u_k ($k = 1, \ldots, N_u$) is available, the state vectors $\vec{x}_i$ can be reconstructed (figure 2), i.e., by the time delay method [31, 32] which generates a state vector of dimension m from a time series of length N_u:

$$\vec{x}_i = \sum_{j=1}^{m} u_{i+\tau(j-1)}\vec{e}_j, \tag{2}$$

where $\vec{e}_j$ is the unit vector with $(\vec{e}_j)_j = 1$. The length of the constructed time sequence $\{\vec{x}_i\}$ is $N = N_u - \tau(m-1)$.

Different kinds of dynamics lead to characteristically different patterns in the corresponding RPs (figure 1(D)). Such differences are quantified with the *RQA* (figures 1(E) and 2). The first RQA measures introduced were mainly based on the distribution of diagonal lines (and their lengths) in the RP, expressed by the length distribution $H(\ell)$ counting the number of diagonal lines in the RP that have exact length ℓ. One interesting, frequently used RQA measure that uses $H(\ell)$ is the *determinism DET*, which is the fraction of recurrence points forming diagonal lines of at least length ℓ in the RP

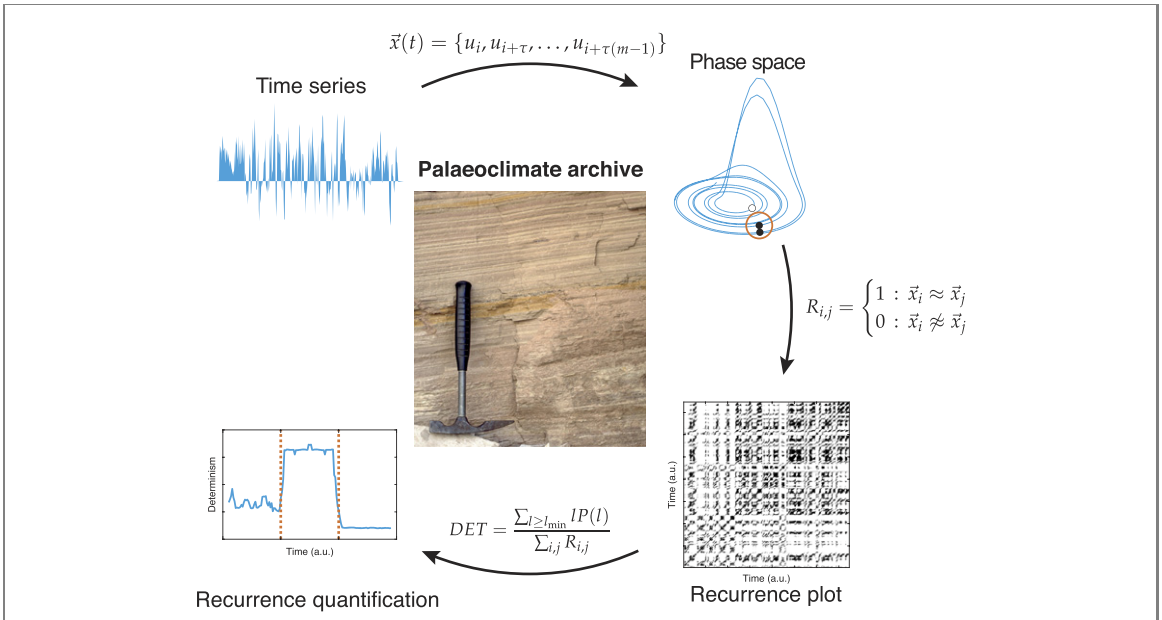

Figure 2. Schematic illustration of RP and RQA. A time series (top left) is first embedded in a higher-dimensional phase space (top right), before the recurrence matrix is computed (bottom right). From the recurrence matrix, measures such as determinism are computed (bottom left), which can then be interpreted in terms of changing dynamics of the underlying system.

$$\mathrm{DET} = \frac{\sum_{\ell \geqslant \ell_{\min}} \ell H(\ell)}{\sum_{i,j} R_{i,j}}. \tag{3}$$

It measures the likelihood that the dynamics of the system sustains to follow a dynamics that had been already occurred at a previous time; it is, therefore, related to the local predictability of the dynamics.

A further important progress was to identify the binary, square matrix **R** as the adjacency matrix of a complex network [33, 34]. The resulting *recurrence network* $A_{i,j} = R_{i,j} - \delta_{i,j}$ consists of nodes representing the time points of the phase space trajectory and links that represent the similarity (recurrence) between a pair of time points. The Kronecker delta $\delta_{i,j}$ is applied to avoid self-loops. The known network measures can be used as additional diagnostic tools for time series analysis that complement the other measures obtained from RPs [33–36].

Besides the application of RP quantification for the classification of different dynamics, studying the variation of a recurrence measure over time is a fundamental application and is successfully used to detect transitions between different dynamical regimes [37–39]. A sliding window is applied to the time series and the recurrence measures are calculated and stored within each window, providing the temporal change of these measures. (Abrupt) changes in the first two statistical moments (mean and variance) are clearly visible in the RP by changes of density of points and block-like pattern. However, more subtle, qualitative changes, such as critical transitions or chaos-chaos transitions, are not directly detectable by the first two statistical moments, but can be identified by changes in the recurrence structure (e.g., line length distribution) and measured by the RQA measures. The fundamental issue when drawing conclusions about the variation of these RP based measures is whether the variation over time is significant or not. In order to get an impression about the significance of the results, a statistical approach is required that provides a confidence interval for the RP measures. The approach depends on the null-hypothesis. For specific null-hypothesis, such as serial independence (corresponding to the vertical line structure in the RP) or stochastic dynamics, a test statistic can be derived analytically [40, 41]. However, dynamical changes are often not as simple and require either surrogate [42, 43] or bootstrap tests [44, 45]. For example, a bootstrap test can use the distribution of line lengths within all sliding windows together to bootstrap new line length distributions out from this merged line length distribution [45].

The transition detection based on recurrence analysis is of high interest for the identification of abrupt or gradual changes and transitions in palaeoclimate dynamics [17, 46–49]. For example, terrestrial, Holocene palaeoclimate records (based on speleothems) from China and Australia have been analysed and their transitions compared. The considered palaeoclimate proxy data represent the northern and the southern extent of the complex East Asian–Indonesian–Austral summer monsoon, where the East Asian summer monsoon and the Indonesian–Austral summer monsoon mutually influence each other. Before analyzing the data, a novel, difference-filter based interpolation schema was applied, because the considered palaeoclimate time series were irregularly sampled and differed in their sampling points [48]. The recurrence analysis has identified periods of

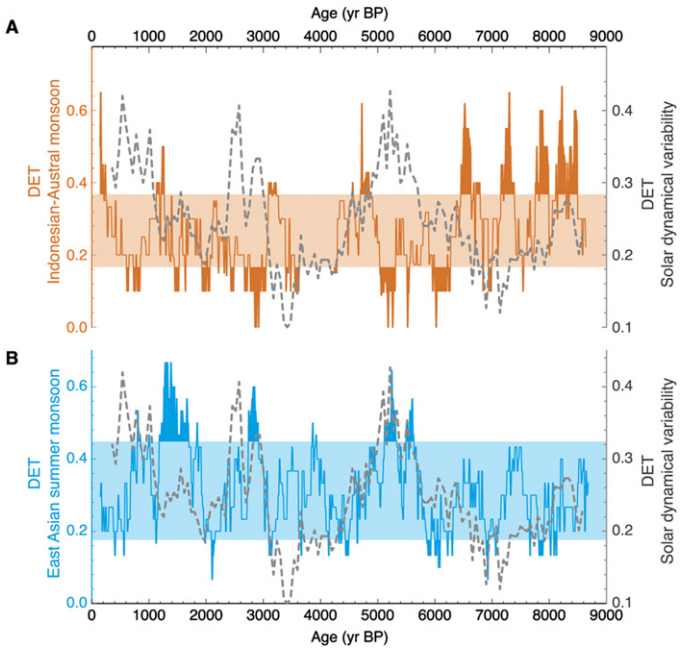

Figure 3. Determinism of palaeoclimate time series of the (A) Indonesian–Austral and (B) East Asian summer monsoon. Statistically significant excursions are indicated by exceeding the horizontal bands. Variations in the solar dynamics are presented by the dashed lines. Figure modified after [50].

alternating, see-saw like weaker and stronger regular dynamics for those palaeoclimate records (figure 3). The variation in regularity in the dynamics can be understood in terms of strong and weak monsoons anti-phased between Asia and Australia [50]. A comparison with variability in solar dynamics suggested that the change from weak to strong monsoon in the East Asian summer monsoon and, vice versa, in the Indonesian–Austral summer monsoon, was probably triggered by solar variations via shifting the position of the intertropical convergence zone [51].

Recurring patterns in data are also often discussed with respect to external drivers, such as the solar variability or Milankovich cycles. It is interesting and helpful to compare the recurrence properties of the external forcing with those of climate data. For this purpose, several approaches for multivariate recurrence analysis are available, e.g., those that study temporal differences [52, 53], different types of synchronization [38, 54], and even causal relationships [55, 56].

By using the novel idea of combining mapograms with RPs, spatio-temporal recurrences of phytoplankton growth with respect to potential external drivers was studied [57]. This study used satellite based variation of chlorophyll concentration in the Southern Californian Bight from 1998 to 2016. RPs for chlorophyll concentration as well as for the sea surface temperature (SST) in the Southern Californian Bight were constructed using mapograms [58]. From the RPs, the potential external forcing were extracted [59] and compared to the El Niño/Southern Oscillation (ENSO) index NINO3.4. in order to better understand the mechanism of phytoplankton growth. A remarkable coincidence between the variation of the NINO3.4 index and the detected driving forces derived from the chlorophyll and SST was found. This supports the hypothesis of an impact of the ENSO on the phytoplankton growth via SST.

In the previous examples, the dynamics of different systems have been compared by a simple comparison of certain recurrence properties. But the RP approach allows even more sophisticated investigations of couplings and synchronization. *Cross RPs* have been introduced to test for the simultaneous occurrence of a similar state in two systems [52, 60]. Besides testing for interrelationship based on similar states, a CRP can be used to visualize temporal divergence between similar dynamical systems [53]. An alternative for testing for similar states in two systems, is to test for the simultaneous occurrence of recurrences in two different systems by *joint RPs*. Joint recurrences are important when looking at generalized synchronizations [54] or coupling directions [61, 62]. For example, using the fraction of recurrence points in the RP and the JRP, we can use concepts from information theory and define conditional measures of dependence [56, 63].

The recurrence measure of dependence [63] has been used to study the feedback mechanism between the Amazonian hydroclimatology and the tropical north Atlantic ocean [64]. The study found that precipitation over the Amazon region controls the atmospheric pressure gradient between the tropical north Atlantic and the Amazonian on short time scales (up to a few weeks). This pressure gradient controls the zonal winds over the

tropical north Atlantic on short and intermediate time scales (up to a year). These zonal winds, finally, influence the tropical Atlantic SST, which, to close the feedback loop, have an impact on precipitation at the Amazon region. These results show that the Amazon region also plays a key role in the tropical Atlantic warming, reinforcing the feedback and triggering severe droughts.

2.2. Future challenges and opportunities for RP analysis

In contrast to graph theory and complex network science, RP based analysis is much younger and is used by a much smaller scientific community. Therefore, many issues are still open and several challenges have to be addressed, enough room for potential theoretical work and improvements.

The scientific communities are becoming more interested in considering uncertainties that come along with the data. New approaches have been suggested, such as simple Monte Carlo based approaches [65] or Bayesian based approaches [66]. First ideas to incorporate uncertainties directly into the RP representation have recently been published [67]. The binary RP is replaced by a probability matrix that states how likely a recurrence occurs at a certain pair of time points. However, the quantification of such probability RP is not easy and still deserves future development.

Another important issue is the impact of embedding and sampling (not only for recurrence analysis but on time series analysis in general). In particular in palaeoclimate research, time series have usually non-equidistant sampling points, due to changes in the sedimentation process, in varying secondary transformations of sediments (e.g., diagenesis), gaps during sedimentation (hiatusses), distortions during sampling retrieval (drilling) or sampling in the lab, etc. The standard procedure is often to interpolate the time series to a common new time axis. A higher temporal resolution than the original sampling can prevail additional recurrence structures such as diagonal lines, where actually such lines would not exist. The increasing number of new high-resolution geological data (e.g., [17]) makes the application of recurrence based analysis interesting, but comes with the risk of misinterpretation due to interpolation or changing sampling times. The limits of interpolation have to be carefully considered or alternative approaches, allowing for the direct application of time series analysis or even RPs to irregularly sampled time series [50, 68], should be applied and further developed.

Recurrence analysis will also benefit from the increasing interest in machine learning. Recurrence patterns could be used as feature vectors for neural network based classification [18, 69] or deep learning based forecasting [70]. Combining recurrence analysis with machine learning concepts will help in investigating the recurrence structures in large sets of big data.

We expect a further increase of the popularity of RP based analysis and its increasing application in further scientific fields, such as turbulence, plasma physics, hydrology, neuroscience, physiology, sociology, etc.

3. Complex networks for climate data analysis and extreme-event prediction

3.1. State-of-the-art

In the last two decades, complex networks have proven to be powerful tools for the quantitative analysis of spatial dependency patterns of measured or simulated climatic observables.

The key concept is that of functional networks, which is also widely used in the analysis of data from physiology or neuroscience. In this approach, network links represent associations or functional dependencies between network nodes, rather than actual physical connections as in the case of anatomical networks. Assume we are given a spatiotemporal dataset $X \in \mathbb{R}^{N\times T}$, where each row $x_i(t) \in \mathbb{R}^T$ denotes a time series encoding the temporal evolution of the climatic variable of interest at a location i. Functional climate networks are then typically constructed in the following way. First, the time series x_i from different geographical locations i are identified with network or graph nodes $v_i \in \mathcal{V}$. Second, statistical dependencies between pairs of time series x_i and x_j are represented by network links $e_{ij} \in \mathcal{E}$ (figure 4).

Time series, e.g., representing temperature, pressure, or rainfall variability, can directly originate from measurement stations [71], from the single cells of spatially gridded data sets [72], or from climate indices representing variability of entire climate modes [73].

To measure the dependencies between time series, a multitude of different similarity measures has been proposed, including Pearson's linear correlation coefficient [29] and modifications thereof [74, 75], nonlinear mutual information [76–78], event synchronization [79, 80] to measure dependencies in highly intermittent rainfall time series, as well as different causality measures [81–83]. Commonly, only the strongest or statistically most significant dependencies are represented by network nodes, while weaker, non-significant dependencies are discarded (figure 5).

The topology of the network or graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ can then be represented by the network adjacency matrix $\mathbf{A}$, for which (in the unweighted and undirected case) $A_{ij} = 1$ if there exists a link e_{ij} between nodes v_i and v_j and $A_{ij} = 0$ otherwise, with straightforward extensions to the cases of directed and weighted networks.

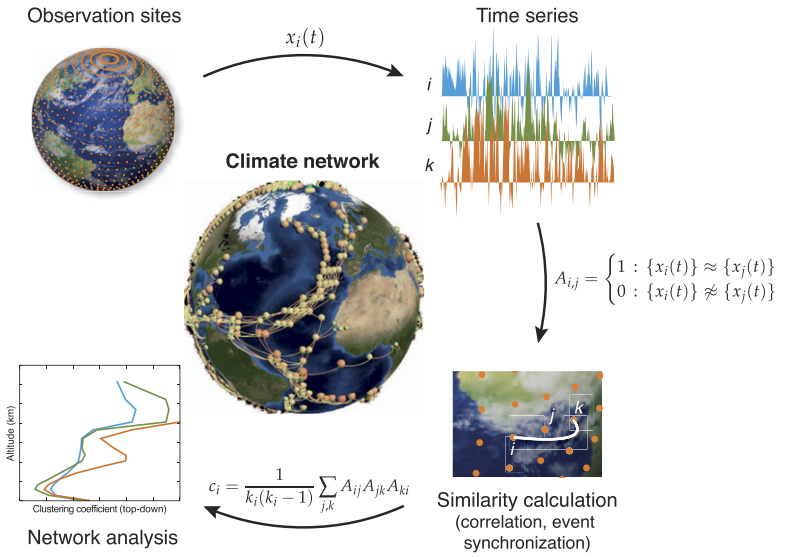

Figure 4. Schematic diagram showing how climate networks are constructed from data. Observation sites (top left) contain time series encoding the variability of a climatic variable of interest (top right), which are identified with the nodes of the network. Links are placed between any two nodes (bottom right) if there is a strong and significant similarity between the corresponding pair of time series. Once this is repeated for all possible pairs of time series, a network is obtained, which encodes the coupling structure of the climatic field of interest, in terms of the network topology. Different aspects of this coupling structure can finally be quantitatively investigated using node- or link-based network measures (bottom left).

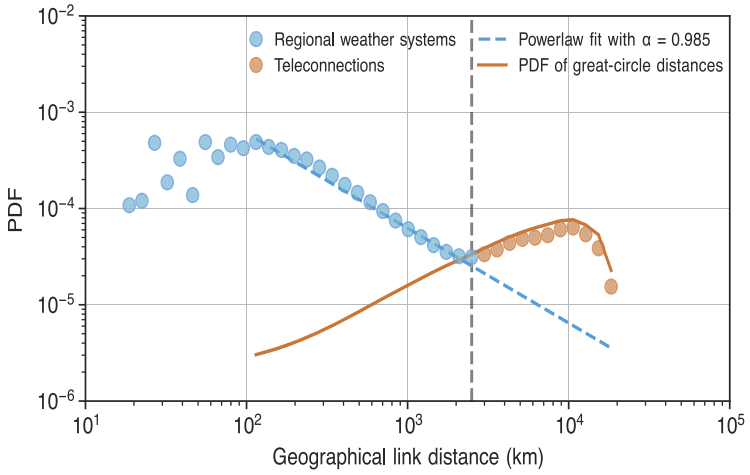

Figure 5. Estimated probability density function of the geographical distances of network links representing significant synchronizations between extreme rainfall events (circles). A break in the scaling around 2500 km is marked by the vertical black dashed line. For distances smaller than this threshold, the PDF approximately follows a power law with exponent close to one (cyan), indicating that the probability of significant extreme-event synchronization scales inversely to the geographical distance. This regime can be associated with regional weather systems. For larger distances, however, the probability of significant extreme-event synchronizations increases again, with peak around 10 000 km. In this super-power law scaling regime, the distance-distribution closely follows the distribution of all possible great circles on the Earth's surface. Atmospheric Rossby waves propagate—to a first approximation—along great circles, and have been suggested to be the main cause for these global-scale teleconnections. Figure adapted from [100].

The key idea behind this functional network construction from observed or simulated data is thus to make the spatial dependency structure of a climatic variable of interest mathematically accessible in terms of the network adjacency matrix **A** (figure 4). In a following step, node-based network measures, quantifying for example different aspects of centrality or clustering, can then be used to investigate specific characteristics of the network topology, and in turn to infer information about the dynamical oceanic or atmospheric processes causing the dependencies. For example, the simple network measure *degree k*, defined at node *i* as

$$k_i = \sum_{j=1}^{N} A_{ij}, \tag{4}$$

quantifies the number of locations that exhibit statistically similar behavior as the time series x_i. Focussing on path structures within the network, the centrality measure *betweenness b* have been shown to be useful in the context of climatic interpretation,

$$b_i = \sum_{j \neq i \neq k} \frac{\sigma_{jk}(i)}{\sigma_{jk}}, \tag{5}$$

where σ_{jk} denotes the number of shortest network paths from k to j and $\sigma_{jk}(i)$ the number of such shortest paths that pass through node i. This measure is expected to be high in regions important for the large-scale information transport in the functional network. Complementarily, the clustering coefficient c

$$c_i = \frac{1}{k_i(k_i - 1)} \sum_{j,k} A_{ij} A_{jk} A_{ki}, \tag{6}$$

measures how strongly links tend to form triangles of links, and thus the tendency toward homogenous, clustered behavior.

It has, for example, been shown in a climate network representing dependencies between extreme rainfall events that high degree and betweenness centrality values correspond to regions that are important for the large-scale propagation of extreme events, whereas high clustering coefficients have been found in regions where large, persistent convective systems frequently develop [84, 85].

Networks thus provide a highly flexible methodological framework for the investigation of spatial covariability patterns in climate dynamics, complementing and extending existing linear methods such as principal component analysis [86]. Generalizations toward interacting networks [87], as well as extensions to multivariate settings [88] and multi-scale approaches [89, 90] have further enhanced the flexibility and broadened the applicability of network-based methods in the climate sciences.

Approaches based on complex networks have been widely employed to study the characteristics and large-scale impacts of the ENSO [91–93]. Focussing on specific correlation characteristics between SST anomalies in the tropical Pacific ocean, it has been shown that networks can be used to predict El Niño events up to one year in advance [94], even crossing the so-called *spring barrier* for El Niño forecasting [95].

Very recently, a bi-variate network approach was implemented to study the influence of tropical Atlantic ocean SSTs on rainfall anomalies in the central Amazon. It was revealed in the latter study that Amazon droughts are preceded by the development of an SST dipole between the northern and the southern tropical Atlantic ocean. This information could be used to establish an early-warning scheme that correctly hindcasts six out of the seven most severe Amazon droughts that occurred during the last four decades [88].

Complex networks have also been extensively used to study spatial synchronization patterns of extreme rainfall events on regional [79, 96–99] and global [90, 100] scales. It has been shown in this context that well-known network centrality and clustering measures can be assigned a climatic interpretation, allowing to associate the spatial patterns exhibited by such network measures with the driving atmospheric mechanisms [80, 87, 99, 101]. For example, our understanding of atmospheric Rossby waves and the mid-latitude jet streams has benefited significantly from analyses employing network-based methods [101, 102]. In particular, it has recently been shown in a climate network analysis that upper-atmospheric Rossby wave trains are the main mechanism causing stable synchronization patterns of extreme rainfall on global scales, leading to a super-power law scaling in the distance distribution of significant synchronizations between extreme events at different locations [100] (figure 3): a power law-scaling was found in the latter study for the geographical distances of network links representing significant synchronizations between extreme rainfall events, if restricted to distances below around 2500 km. For longer distances, however, the distribution strongly deviates from the power law, with much higher probability than an extrapolation from the smaller distances would suggest. This exceptional kind of scaling behavior has been associated with the presence of so-called *dragon kings* [103, 104].

Apart from climatic analysis and identification of atmospheric mechanisms, networks have also been used to evaluate the quality of different reanalysis datasets and the simulations of regional and global climate models concerning the representation of extreme events and their dependency structures [105].

In some cases, network-based approaches have revealed previously unrecognized forecast potential for extreme rainfall events [85, 100, 106]. The key idea in this context is that recurrent, temporally ordered event patterns are captured in the network topology. A great advantage of network-based methods for spatial pattern identification in this context is their flexibility. In particular, they allow to focus the analysis on the dependencies of extreme events alone rather than on lower-order statistical moments as with more traditional methods based on principal component analysis of the covariance matrix. Examples of spatial propagation patterns revealed using complex networks include propagating mesoscale convective systems, frontal systems, or tropical cyclones [85, 106], but also time-delayed synchronization patterns between extremes across large spatial distances caused by quasi-stationary Rossby waves [100]. Assuming stationarity to some degree, such

statistically inferred, time-lagged synchronization patterns can then be used to establish empirically-based forecast rules. Based on insights into spatial dependency patterns gained from network analysis of climate variables over the Indian subcontinent, a competitive prediction scheme for the onset and withdrawal of the Indian summer monsoon has furthermore been established [107].

3.2. Future challenges and opportunities for network-based climate analysis

There is tremendous potential for further developing network-based approaches to investigate spatial co-variability patterns of climatic observables beyond the linear regime. Upon identifying time intervals during which specific, network-derived co-variability structures of interest are active, composites anomalies of confounding climatic variables can be derived to reveal the underlying mechanisms in the circulation dynamics of the atmosphere or oceans.

Also for predictive purposes, it is important to go beyond purely statistical information on dependencies between climatic variables at different locations, in particular to circumvent potential problems induced by non-stationarity. Along the lines of the above, network-based approaches can play an interface role here, connecting statistical information on recurring spatial co-variability patterns with the underlying physical mechanisms.

A particular opportunity for future applications of networks in climate science, which has so far only been touched upon in very few studies, is to compare network-derived spatial dependency patterns from observations with corresponding patterns obtained from simulations by general circulation models. Focussing first on historical simulation runs, e.g., from the coupled model intercomparision project (CMIP), this can allow to evaluate and compare the performance of different state-of-the-art general circulation models with respect to reproducing observed spatial dependency structures. In a next step, future projections from the CMIP models could be analyzed concerning such dependency structures, whilst taking into account the information on their performance in reproducing structures from historical observations.

The recent success of machine and in particular deep learning methods in extracting information from large sets of time series [108] suggests that these techniques also carry great potential for applications in the analysis of climate data [109]. We see great potential in combining methods from complexity science that have already been shown to be applicable for the analysis of geoscientific data—such as complex networks—with state-of-the-art machine learning methods such as recurrent neural networks [110] for the analysis of time series and convolutional neural networks [111] for the analysis and inference of maps between spatial patterns. In particular, such deep learning architectures should prove extremely valuable in systematically extracting the predictive skill that is encoded in the topology of networks representing the (temporally ordered) dependency structure of climate variables at different locations.

Moreover, combinations of network approaches with machine learning techniques promise to be extremely valuable in quantifying the interactions between different parts of the coupled Earth system, such as climate-vegetation interactions. Even more generally, innovative combinations of these methodological frameworks should be capable of improving our understanding of interactions between natural and socioeconomic systems; for example in the context of human migration causes by anthropogenic climate change. We expect combinations of network theory and similar frameworks from complexity science with genuine machine learning approaches to also provide new and valuable insights in other fields where complex natural systems can be studied from a data-centric perspective, such as epidemiology and neuroscience.

4. Conclusion

The problems and challenges arising from the analysis of data obtained from both simulated and real-world complex systems will keep stimulating methodological developments in complexity science and beyond in the decades to come. In addition to the natural complex systems that humans have started to study already millennia before today, we are now producing ever more complex systems ourselves—such as the internet, social networks, economic, or financial systems.

A key driver of further advances is the desire to improve predictions of the behaviour of complex systems and especially—for example in the context of the ongoing climate change driven by the anthropogenic release of greenhouse gases—of the response of complex systems to time-varying external forcing. But for this purpose, it will be vital to improve our abilities to reveal the structural characteristics of complex systems from data, with focus on both their temporal and spatial intricacies. Advancing our knowledge in this direction will necessarily have to rely on improvements of our capabilities regarding data-driven inference of governing principles, in order to reach a deeper understanding of the connection between the microscopic dynamics of complex system constituents and their nonlinear interactions on the one hand, and the dynamics emerging from these interactions at the macroscopic level, on the other hand. We are convinced that—among many

other branches of complexity science—recurrence analysis and complex networks will have a crucial role to play in this endeavour.

Data availability

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

Acknowledgments

We thank our students and collaborators for fruitful discussions and performing some of the mentioned analysis, in particular Deniz Eroglu for performing the study on the Indonesian–Austral and East Asian summer monsoon. We acknowledge the financial support by the Volkswagen Foundation, DFG Project No. MA4759/9, the European Union's Horizon 2020 research and innovation program under Grant Agreement No. 820970 (TiPES #72), the BMBF project climXtreme, and the Russian Ministry of Science and Education Agreement (No. 13.1902.21.0026).

ORCID iDs

Niklas Boers https://orcid.org/0000-0002-1239-9034

References

[1] Hawkins G S 1964 *Nature* **202** 1258–61
[2] Newton R R and Jenkins R E 1972 *Nature* **239** 511–2
[3] Pedersen O 2009 *Early Physics and Astronomy: A Historical Introduction* (Cambridge: Cambridge University Press)
[4] Poincaré H 1890 *Acta Math.* **13** 1–271
[5] Carathéodory C 1919 *Sitzungsberichte der Preussischen Akademie der Wissenschaften* 24 (Berlin: Berlin-Brandenburgische Akademie der Wissenschaften) pp 580–4
[6] Elton C and Nicholson M 1942 *J. Anim. Ecol.* **11** 215
[7] Schuster A 1898 *J. Geophys. Res.* **3** 13
[8] Schuster A 1906 *Phil. Trans. R. Soc.* A **206** 69–100
[9] Yule G U 1927 *Phil. Trans. R. Soc.* A **226** 267–98
[10] Box G E P and Jenkins G M 1976 *Time Series Analysis* (San Francisco, CA: Holden-Day)
[11] Priestley M B 1981 *Spectral Analysis and Time Series* (London: Academic)
[12] Kantz H and Schreiber T 1997 *Nonlinear Time Series Analysis* (Cambridge: University Press)
[13] Eckmann J-P, Kamphorst S O and Ruelle D 1987 *Europhys. Lett.* **4** 973–7
[14] Zbilut J P and Webber C L Jr 1992 *Phys. Lett.* A **171** 199–203
[15] Hou Y, Aldrich C, Lepkova K, Machuca L L and Kinsella B 2016 *Corros. Sci.* **112** 63–72
[16] Ngamga E J, Bialonski S, Marwan N, Kurths J, Geier C and Lehnertz K 2016 *Phys. Lett.* A **380** 1419–25
[17] Westerhold T *et al* 2020 *Science* **369** 1383–7
[18] Afonso L C S, Rosa G H, Pereira C R, Weber S A T, Hook C, Albuquerque V H C and Papa J P 2019 *Future Gener. Comput. Syst.* **94** 282–92
[19] Donges J F, Donner R V, Rehfeld K, Marwan N, Trauth M H and Kurths J 2011 *Nonlinear Process Geophys.* **18** 545–62
[20] Pawar S A, Seshadri A, Unni V R and Sujith R I 2017 *J. Fluid Mech.* **827** 664–93
[21] Costa L d F, Rodrigues F A, Travieso G and Villas Boas P R 2007 *Adv. Phys.* **56** 167–242
[22] Lacasa L, Luque B, Ballesteros F, Luque J and Nuño J C 2008 *Proc. Natl Acad. Sci. USA* **105** 4972–5
[23] Donges J F, Zou Y, Marwan N and Kurths J 2009 *Europhys. Lett.* **87** 48007
[24] Rubinov M and Sporns O 2010 *NeuroImage* **52** 1059–69
[25] Campanharo A S L O, Sirer M I, Malmgren R D, Ramos F M and Amaral L A N 2011 *PloS One* **6** e23378
[26] Schultz P, Heitzig J and Kurths J 2014 *New J. Phys.* **16** 125001
[27] Gao J, Barzel B and Barabási A-L 2016 *Nature* **530** 307–12
[28] Braga A C, Alves L G A, Costa L S, Ribeiro A A, de Jesus M M A, Tateishi A A and Ribeiro H V 2016 *Physica* A **444** 1003–11
[29] Tsonis A A and Roebber P J 2004 *Physica* A **333** 497–504
[30] Marwan N, Wessel N, Meyerfeldt U, Schirdewan A and Kurths J 2002 *Phys. Rev.* E **66** 026702
[31] Takens F 1981 Detecting strange attractors in turbulence *Dynamical Systems and Turbulence* (*Lecture Notes in Mathematics* vol 898) ed D Rand and L S Young (Berlin: Springer) pp 366–81
[32] Packard N H, Crutchfield J P, Farmer J D and Shaw R S 1980 *Phys. Rev. Lett.* **45** 712–6
[33] Marwan N, Donges J F, Zou Y, Donner R V and Kurths J 2009 *Phys. Lett.* A **373** 4246–54
[34] Zou Y, Donner R V, Marwan N, Donges J F and Kurths J 2019 *Phys. Rep.* **787** 1–97
[35] Donner R V, Zou Y, Donges J F, Marwan N and Kurths J 2010 *New J. Phys.* **12** 033025
[36] Donner R V, Small M, Donges J F, Marwan N, Zou Y, Xiang R and Kurths J 2011 *Int. J. Bifurcation Chaos* **21** 1019–46
[37] Trulla L L, Giuliani A, Zbilut J P and Webber C L Jr 1996 *Phys. Lett.* A **223** 255–60
[38] Marwan N, Carmen Romano M, Thiel M and Kurths J 2007 *Phys. Rep.* **438** 237–329
[39] Marwan N 2011 *Int. J. Bifurcation Chaos* **21** 1003–17
[40] Aparicio T, Pozo E F and Saura D 2008 *J. Econ. Behav. Organ.* **65** 768–87
[41] Hirata Y and Aihara K 2011 *Int. J. Bifurcation Chaos* **21** 1077–84
[42] Thiel M, Romano M C, Kurths J, Rolfs M and Kliegl R 2006 *Europhys. Lett.* **75** 535–41

[43] Lancaster G, Iatsenko D, Pidde A, Ticcinelli V and Stefanovska A 2018 *Phys. Rep.* **748** 1–60
[44] Schinkel S, Marwan N, Dimigen O and Kurths J 2009 *Phys. Lett.* A **373** 2245–50
[45] Marwan N, Schinkel S and Kurths J 2013 *Europhys. Lett.* **101** 20007
[46] Donges J F, Donner R V, Trauth M H, Marwan N, Schellnhuber H-J and Kurths J 2011 *Proc. Natl Acad. Sci.* **108** 20422–7
[47] Marwan N and Kurths J 2015 *Chaos* **25** 097609
[48] Ozken I, Eroglu D, Stemler T, Marwan N, Bagci G B and Kurths J 2015 *Phys. Rev.* E **91** 062911
[49] Trauth M H, Asrat A, Duesing W, Foerster V, Kraemer K H, Marwan N, Maslin M A and Schaebitz F 2019 *Clim. Dyn.* **53** 2557–72
[50] Eroglu D, McRobie F H, Ozken I, Stemler T, Wyrwoll K H, Breitenbach S F M, Marwan N and Kurths J 2016 *Nat. Commun.* **7** 12929
[51] Wang Y *et al* 2005 *Science* **308** 854–7
[52] Marwan N, Thiel M and Nowaczyk N R 2002 *Nonlinear Process Geophys.* **9** 325–31
[53] Marwan N and Kurths J 2005 *Phys. Lett.* A **336** 349–57
[54] Romano M C, Thiel M, Kurths J and von Bloh W 2004 *Phys. Lett.* A **330** 214–23
[55] Feldhoff J H, Donner R V, Donges J F, Marwan N and Kurths J 2012 *Phys. Lett.* A **376** 3504–13
[56] Ramos A M T, Builes-Jaramillo A, Poveda G, Goswami B, Macau E E N, Kurths J and Marwan N 2017 *Phys. Rev.* E **95** 052206
[57] Riedl M, Marwan N and Kurths J 2017 *Eur. Phys. J.: Spec. Top.* **226** 3273–85
[58] Riedl M, Marwan N and Kurths J 2015 *Chaos* **25** 123111
[59] Casdagli M C 1997 *Physica* D **108** 12–44
[60] Marwan N and Kurths J 2002 *Phys. Lett.* A **302** 299–307
[61] Romano M C, Thiel M, Kurths J and Grebogi C 2007 *Phys. Rev.* E **76** 036211
[62] Zou Y, Romano M C, Thiel M, Marwan N and Kurths J 2011 *Int. J. Bifurcation Chaos* **21** 1099–111
[63] Goswami B, Marwan N, Feulner G and Kurths J 2013 *Eur. Phys. J.: Spec. Top.* **222** 861–73
[64] Builes-Jaramillo A, Marwan N, Poveda G and Kurths J 2018 *Clim. Dyn.* **50** 2951–69
[65] Breitenbach S F M *et al* 2012 *Clim. Past* **8** 1765–79
[66] Schütz N and Holschneider M 2011 *Phys. Rev.* E **84** 021120
[67] Goswami B, Boers N, Rheinwalt A, Marwan N, Heitzig J, Breitenbach S F M and Kurths J 2018 *Nat. Commun.* **9** 48
[68] Ozken I, Eroglu D, Breitenbach S F M, Marwan N, Tan L, Tirnakli U and Kurths J 2018 *Phys. Rev.* E **98** 052215
[69] Hatami N, Gavet Y and Debayle J 2018 *Proc. SPIE* **10696** 106960Y
[70] Estebsari A and Rajabi R 2020 *Electronics* **9** 68
[71] Rheinwalt A, Marwan N, Kurths J, Werner P and Gerstengarbe F-W 2012 *Europhys. Lett.* **100** 28002
[72] Donges J F, Zou Y, Marwan N and Kurths J 2009 *Europhys. Lett.* **87** 48007
[73] Tsonis A A 2007 *Int. J. Bifurcation Chaos* **17** 4229–43
[74] Yamasaki K, Gozolchiani A and Havlin S 2008 *Phys. Rev. Lett.* **100** 228501
[75] Ciemer C, Boers N, Barbosa H M J, Kurths J and Rammig A 2018 *Clim. Dyn.* **51** 371–82
[76] Donges J F, Zou Y, Marwan N and Kurths J 2009 *Eur. Phys. J.: Spec. Top.* **174** 157–79
[77] Barreiro M, Marti A C and Masoller C 2011 *Chaos* **21** 013101
[78] Deza J I, Barreiro M and Masoller C 2015 *Chaos* **25** 033105
[79] Malik N, Bookhagen B, Marwan N and Kurths J 2012 *Clim. Dyn.* **39** 971–87
[80] Boers N, Bookhagen B, Marwan N, Kurths J and Marengo J 2013 *Geophys. Res. Lett.* **40** 4386–92
[81] Hlinka J, Hartman D, Vejmelka M, Runge J, Marwan N, Kurths J and Paluš M 2013 *Entropy* **15** 2023–45
[82] Runge J *et al* 2015 *Nat. Commun.* **6** 8502
[83] Runge J, Nowack P, Kretschmer M, Flaxman S and Sejdinovic D 2019 *Sci. Adv.* **5** eaau4996
[84] Boers N, Bookhagen B, Marwan N, Kurths J and Marengo J 2013 *Geophys. Res. Lett.* **40** 4386–92
[85] Boers N, Rheinwalt A, Bookhagen B, Barbosa H M J, Marwan N, Marengo J and Kurths J 2014 *Geophys. Res. Lett.* **41** 7397–405
[86] Donges J F, Petrova I, Loew A, Marwan N and Kurths J 2015 *Clim. Dyn.* **45** 2407–24
[87] Donges J F, Schultz H C H, Marwan N, Zou Y and Kurths J 2011 *Eur. Phys. J.* B **84** 635–51
[88] Ciemer C, Rehm L, Kurths J, Donner R V, Winkelmann R and Boers N 2020 *Environ. Res. Lett.* **15** 094087
[89] Agarwal A, Marwan N, Rathinasamy M, Merz B and Kurths J 2017 *Nonlinear Process Geophys.* **24** 599–611
[90] Agarwal A, Caesar L, Marwan N, Maheswaran R, Merz B and Kurths J 2019 *Sci. Rep.* **9** 8808
[91] Gozolchiani A, Havlin S and Yamasaki K 2011 *Phys. Rev. Lett.* **107** 148501
[92] Wiedermann M, Radebach A, Donges J F, Kurths J and Donner R V 2016 *Geophys. Res. Lett.* **43** 7176–85
[93] Fan J, Meng J, Ashkenazy Y, Havlin S and Schellnhuber H J 2017 *Proc. Natl Acad. Sci.* **114** 201701214
[94] Ludescher J, Gozolchiani A, Bogachev M I, Bunde A, Havlin S and Schellnhuber H J 2013 *Proc. Natl Acad. Sci.* **110** 11742–5
[95] Meng J, Fan J, Ludescher J, Agarwal A, Chen X, Bunde A, Kurths J and Schellnhuber H J 2020 *Proc. Natl Acad. Sci. USA* **117** 177–83
[96] Stolbova V, Martin P, Bookhagen B, Marwan N and Kurths J 2014 *Nonlinear Process Geophys.* **21** 901–17
[97] Rheinwalt A, Boers N, Marwan N, Kurths J, Hoffmann P, Gerstengarbe F W and Werner P 2016 *Clim. Dyn.* **46** 1066–74
[98] Agarwal A, Marwan N, Maheswaran R, Merz B and Kurths J 2018 *J. Hydrol.* **563** 802–10
[99] Kurths J, Agarwal A, Shukla R, Marwan N, Rathinasamy M, Caesar L, Krishnan R and Merz B 2019 *Nonlinear Process Geophys.* **26** 251–66
[100] Boers N, Goswami B, Rheinwalt A, Bookhagen B, Hoskins B and Kurths J 2019 *Nature* **566** 373–7
[101] Boers N, Bookhagen B, Barbosa H M J, Marwan N, Kurths J and Marengo J A 2014 *Nat. Commun.* **5** 5199
[102] Wang Y, Gozolchiani A, Ashkenazy Y, Berezin Y, Guez O and Havlin S 2013 *Phys. Rev. Lett.* **111** 138501
[103] Sornette D 2009 Dragon-Kings, Black Swans and the Prediction of Crises *Swiss Finance Institute Research Paper No. 09-36,* Available at SSRN: https://ssrn.com/abstract=1470006 or http://dx.doi.org/10.2139/ssrn.1470006
[104] Peters O, Christensen K and Neelin J D 2012 *Eur. Phys. J.: Spec. Top.* **205** 147–58
[105] Boers N, Bookhagen B, Marengo J, Marwan N, von Storch J-S and Kurths J 2015 *J. Clim.* **28** 1031–56
[106] Boers N, Bookhagen B, Marwan N and Kurths J 2016 *Clim. Dyn.* **46** 601–17
[107] Stolbova V, Surovyatkina E, Bookhagen B and Kurths J 2016 *Geophys. Res. Lett.* **43** 3982–90
[108] Fulcher B D and Jones N S 2017 *Cell Syst.* **5** 527–31
[109] Reichstein M, Camps-Valls G, Stevens B, Jung M, Denzler J, Carvalhais N and Prabhat 2019 *Nature* **566** 195–204
[110] Connor J T, Martin R D and Atlas L E 1994 *IEEE Trans. Neural Netw.* **5** 240–54
[111] Fukushima K 1980 *Biol. Cybern.* **36** 193–202