Stoch Environ Res Risk Assess (2014) 28:1685–1708
DOI 10.1007/s00477-013-0825-8

ORIGINAL PAPER

Complexity–entropy analysis of daily stream flow time series in the continental United States

Francesco Serinaldi · Luciano Zunino · Osvaldo A. Rosso

Published online: 27 November 2013

Abstract Complexity–entropy causality plane (CECP) is a diagnostic diagram plotting normalized Shannon entropy $\mathcal{H}_S$ versus Jensen–Shannon complexity $\mathcal{C}_{JS}$ that has been introduced in nonlinear dynamics analysis to classify signals according to their degrees of randomness and complexity. In this study, we explore the applicability of CECP in hydrological studies by analyzing 80 daily stream flow time series recorded in the continental United States during a period of 75 years, surrogate sequences simulated by autoregressive models (with independent or long-range memory innovations), Theiler amplitude adjusted Fourier transform and Theiler phase randomization, and a set of signals drawn from nonlinear dynamic systems. The effect of seasonality, and the relationships between the CECP quantifiers and several physical and statistical properties of the observed time series are also studied. The results point out that: (1) the CECP can discriminate chaotic and stochastic signals in presence of moderate observational noise; (2) the signal classification depends on the sampling frequency and aggregation time scales; (3) both chaotic and stochastic systems can be compatible with the daily stream flow dynamics, when the focus is on the information content, thus setting these results in the context of the debate on observational equivalence; (4) the empirical relationships between $\mathcal{H}_S$ and $\mathcal{C}_{JS}$ and Hurst parameter H, base flow index, basin drainage area and stream flow quantiles highlight that the CECP quantifiers can be considered as proxies of the long-term low-frequency groundwater processes rather than proxies of the short-term high-frequency surface processes; (6) the joint application of linear and nonlinear diagnostics allows for a more comprehensive characterization of the stream flow time series.

Keywords Stream flow · Complexity–entropy causality plane · Permutation entropy · Permutation statistical complexity · Bandt and Pompe method · Hurst parameter

F. Serinaldi (✉)
School of Civil Engineering and Geosciences, Newcastle University, Newcastle Upon Tyne NE1 7RU, UK
e-mail: francesco.serinaldi@newcastle.ac.uk

F. Serinaldi
Willis Research Network, 51 Lime St., London EC3M 7DQ, UK

L. Zunino
Centro de Investigaciones Ópticas (CONICET La Plata-CIC), C.C. 3, 1897 Gonnet, Argentina
e-mail: lucianoz@ciop.unlp.edu.ar

L. Zunino
Departamento de Ciencias Básicas, Facultad de Ingeniería, Universidad Nacional de La Plata (UNLP), 1900 La Plata, Argentina

O. A. Rosso
Laboratorio de Sistemas Complejos, Facultad de Ingeniería, Universidad de Buenos Aires, 1063 Av. Paseo Colón 840, Ciudad Autónoma de Buenos Aires, Argentina
e-mail: oarosso@gmail.com

O. A. Rosso
Instituto de Física, Universidade Federal de Alagoas, BR 104 Norte km 97, Maceió, AL 57072-970, Brazil

1 Introduction

The need for classification of catchments and hydrological phenomena has received an increasing attention in the literature in the last years (McDonnell and Woods 2004; Sivakumar et al. 2007; Sivakumar and Singh 2012) owing to the importance of making an appropriate identification of models' type and complexity and data requirements. A

suitable classification framework can allow recognizing the dominant processes of the hydrological systems, leading to an overall simplification of the data collection (Grayson and Blöschl 2000; Sivakumar et al. 2007) and the use of modeling strategies that are in middle course between the treatment of all hydrological systems in the same way, despite their differences, and the treatment of each system as a unique one, despite the systems' similarities (Sivakumar and Singh 2012). In this context, classification plays a key role to transfer the information from gauged to ungauged basins for modeling and prediction (e.g., Castellarin et al. 2001).

Looking at the stream flow generating mechanism from the nonlinear dynamic viewpoint, Sivakumar (2004) proposed a classification method based on the concepts of "extent of complexity" or "dimensionality", which has been further developed by Sivakumar et al. (2007), introducing a simple phase space reconstruction technique. Sivakumar and Singh (2012) further highlighted that system complexity is an appropriate basis for the classification framework and nonlinear dynamic concepts constitute a suitable methodology for assessing system complexity.

However, the concepts of "system" and "complexity" are not uniquely defined in the literature. A reasonable definition of system proposed by Dooge (1968) in a hydrologic context and adopted by Sivakumar and Singh (2012) is "Any structure, device, scheme, or procedure, real or abstract, that interrelates in a given time reference, an input, cause, or stimulus, of matter, energy, or information, and an output, effect, or response, of information, energy, or matter". On the other hand, many definitions of complexity are available, especially from physics and information theory (López-Ruiz et al. 2011); for example, Kolmogorov–Chaitin algorithmic complexity (Kolmogorov 1965; Chaitin 1966), the Lempel–Ziv complexity (Lempel and Ziv 1976), the effective measure complexity of Grassberger ($\mathcal{C}_E$; Grassberger 1986), the ε-machine complexity (Crutchfield and Young 1989), among others.

The choice of a specific definition of complexity and therefore the study of a system must be performed according to the objective of the inquiry which can be specified for instance in terms of process, scale, and purpose of interest (Sivakumar 2008; Sivakumar and Singh 2012). Sivakumar and Singh (2012) assessed the complexity of a system in terms of variability of the data through dimension estimation. Krasovskaia (1995) and Krasovskaia (1997) suggested a quantitative approach of river flow regime classification based on Shannon entropy S and the minimization of an entropy-based objective function. Hauhs and Lange (2008) introduced symbolic dynamics to quantify randomness and complexity of stream flow time series by applying two information quantifiers: the mean information gain $\mathcal{H}_G$, which measures the randomness of a time series in a nonlinear manner, and the fluctuation complexity (σ_λ^2), which is related to the structure of conditional probabilities. Pan et al. (2011) applied $S, \mathcal{H}_G, \mathcal{C}_E$ and σ_λ^2 to compare patterns in observed and simulated time series of soil moisture contents and corresponding rainfall, confirming previous results of Pachepsky et al. (2006). Pan et al. (2012) used the same information metrics to characterize 5-year stream flow and precipitation time series data from two monitored agricultural watersheds managed by the United States Department of Agriculture (USDA). These studies rely on diagnostic diagrams plotting entropy and complexity measures (e.g., σ_λ^2 vs. $\mathcal{H}_G$ and $\mathcal{H}_E$ vs. S) to characterize the temporal structure of the analyzed signals (at single or multiple time scales), trying to distinguish rainfall, soil water content and stream flow processes in terms of information content.

In this study, we investigate the applicability of a diagnostic plot based on the concepts of Shannon entropy, Jensen–Shannon divergence and phase space reconstruction called complexity–entropy causality plane (CECP), which has been proposed by Rosso et al. (2007a) for examining the properties of a signal. The method merges entropy and nonlinear dynamic concepts, is fully non-parametric (as it is rank-based and does not depend on the shape of the distribution function of the data) and focuses on the temporal structure (complexity) of the signal through an appropriate definition of the probabilities involved in the entropy definition. We explore the capability of the CECP to discriminate the stream flow dynamics in terms of complexity (defined in the next section) and its suitability as a complementary graphical tool for stream flow analysis. The CECP is studied analyzing synthetic time series drawn from stochastic and chaotic systems and 80 daily stream flow time series with 75 years of observations recorded in the continental United States (US). The heterogeneous nature of these 80 stream flow records (described in the next sections) makes this dataset a suitable test bed for exploring CECP applicability in hydrological analyses. The remainder of this paper is organized as follows. In Sects. 2 and 3, the methodology is described in detail resorting to synthetic time series to illustrate the CECP rationale and its performance for signals that exhibit prescribed properties. In Sect. 4, we provide a detailed description of the dataset and a preliminary analysis based on linear methods such as autocorrelation function and power spectrum. Moreover, the results of CECP analysis are shown along with an examination of the empirical relationships between CECP quantifiers (see later for definitions) and several properties of the stream flow series. The effect of the seasonality on CECP is also investigated. Discussion, conclusions and future perspectives are reported in Sects. 5 and 6.

2 Randomness and complexity quantifiers from permutation information theory

2.1 Shannon entropy and statistical complexity

The Shannon entropy is often used as the first natural entropy measure. Focusing on the discrete version, given any arbitrary probability distribution $P = \{p_i: i = 1, \ldots, M\}$, the widely known Shannon logarithmic information measure, $S[P] = -\sum_{i=1}^{M} p_i \log(p_i)$, is related to the information associated with the physical process described by P. If $S[P] = 0$ the knowledge of the underlying process described by the probability distribution is maximal and the possible outcomes can be predicted with complete certainty. On the other hand, our knowledge is minimal for a uniform distribution since every outcome exhibits the same probability of occurrence. The perfect crystal and the isolated ideal gas are two typical examples of systems with minimum and maximum entropy, respectively. However, they are also examples of simple models and therefore of systems with zero complexity, as the structure of the perfect crystal is completely described by minimal information (i.e., distances and symmetries that define the elementary cell) and the probability distribution for the accessible states is centered around a prevailing state of perfect symmetry. On the other hand, all the accessible states of the ideal gas occur with the same probability and can be described by a "simple" uniform distribution. According to López-Ruiz et al. (2011) and using an oxymoron, an object, a procedure, or system is said to be complex when it does not exhibit patterns regarded as simple. It follows that a suitable complexity measure should vanish both for completely ordered and for completely random systems and cannot only rely on the concept of information (which is maximal and minimal for the above mentioned systems).

A suitable measure of complexity can be defined as the product of a measure of information and a measure of disequilibrium, i.e. some kind of distance from the equiprobable distribution of the accessible states of a system. In this respect, Lamberti et al. (2004) introduced an effective statistical complexity measure (SCM) that is able to detect essential details of the dynamics. This SCM is defined as (López-Ruiz et al. 1995)

$$\mathcal{C}_{JS}[P] := \mathcal{Q}_J[P, P_e]\mathcal{H}_S[P] \tag{1}$$

where

$$\mathcal{H}_S[P] := S[P]/S_{\max} \tag{2}$$

is the normalized Shannon entropy ($\mathcal{H}_S \in [0, 1]$) with $S_{\max} = S[P_e] = \log(M)$, $P_e = \{1/M, \ldots, 1/M\}$ is the uniform distribution, and the disequilibrium $\mathcal{Q}_J$ is defined in terms of the extensive (in the thermodynamical sense) Jensen–Shannon divergence. Namely, $\mathcal{Q}_J[P, P_e] = \mathcal{Q}_0\mathcal{J}[P, P_e]$ with $\mathcal{J}[P, P_e] = S[(P + P_e)/2] - S[P]/2 - S[P_e]/2$ and $\mathcal{Q}_0$ is a normalization constant equal to the inverse of the maximum possible value of $\mathcal{J}[P, P_e]$, so that $\mathcal{Q}_J \in [0, 1]$. $\mathcal{Q}_0$ is obtained when one of the components of P, say p_m, is equal to one and the remaining p_i are equal to zero. The Jensen–Shannon divergence quantifies the difference between two (or more) probability distributions and is well-suited to compare the symbol composition between different sequences (Grosse et al. 2002). The $\mathcal{C}_{JS}$ value of a system is null and void in the opposite extreme situations of perfect knowledge (perfect crystal) and maximal randomness (ideal gas), whereas wide range of possible degrees of physical structure does exist between these extreme configurations.

The SCM in Eq. 1 is not a trivial function of the entropy because it measures the interplay between the information stored by the system and the distance from equipartition (measure of a probabilistic hierarchy between the observed parts) of the probability distribution of its accessible states (López-Ruiz et al. 2011). Furthermore, a range of possible SCM values does exist for a given $\mathcal{H}_S$ value (as is shown in Sect. 2.2), meaning that additional information related to the dependence structure between the components of the system and the emergence of nontrivial collective behavior is provided by evaluating the statistical complexity (Martín et al. 2003; Sánchez and López-Ruiz 2005; Escalona-Morán et al. 2010). Moreover, it should be noted that $\mathcal{C}_{JS}$ fulfills two additional properties required for a suitable definition of complexity in physics: (1) the quantifier must be measurable in different physical systems and (2) it should allow for physical interpretation and comparison between two measurements (López-Ruiz et al. 2011). Of course, as mentioned in Sect. 1, other complexity measures can be considered (see Wackerbauer et al. (1994) for a comparison).

Following Zunino et al. (2010b), we apply the statistical complexity for the analysis of time series in order to capture the property of temporal structure of the signals. In this context, we must recall that complexity cannot be univocally measured as it depends on the nature of the description (which always involves a reductionist process) and on the (spatio-temporal) scale of observation (López-Ruiz et al. 2011). Indeed, the definition of complexity in Eq. 1 also depends on the scale. For a given system at each scale of observation, a new set of accessible states appears with its corresponding probability distribution so that complexity changes and therefore different values for $\mathcal{H}_S$ and $\mathcal{C}_{JS}$ are obtained. This aspect is fundamental in geophysical time series analysis (Koutsoyiannis 2010) and will be further investigated in Sect. 3.2

2.2 Bandt–Pompe symbolization method

The evaluation of $\mathcal{H}_S$ and $\mathcal{C}_{JS}$ requires the preliminary definition of a probability distribution P associated with the time series. Bandt and Pompe (2002) introduced a simple

method to define this probability distribution taking into account the time causality of the process dynamics. This approach is based on the symbol sequences that arise naturally from the time series, replacing the observed series with the sequence of the corresponding ranks. Namely, given a time series $\{x_t\colon\ t = 1, \ldots, N\}$, an embedding dimension $D \geq 2\,(D \in \mathbb{N})$, and an embedding delay time $\tau\,(\tau \in \mathbb{N})$, the ordinal pattern of order D generated by

$$s \mapsto \left\{x_{s-(D-1)\tau}, x_{s-(D-2)\tau}, \ldots, x_{s-\tau}, x_s\right\} \tag{3}$$

is considered. For each time instant s, we assign a D-dimensional vector that results from the evaluation of the time series at times $s - (D - 1)\tau, \ldots, s - \tau, s$. Clearly, the higher the value of D, the more information about the past is incorporated into the ensuing vectors. By the ordinal pattern of order D related to the time instant s we mean the permutation $\pi = \{r_0, r_1, \ldots, r_{D-1}\}$ of $\{0, 1, \ldots, D - 1\}$ defined by

$$x_{s-r_0\tau} \geq x_{s-r_1\tau} \geq \cdots \geq x_{s-r_{D-2}\tau} \geq x_{s-r_{D-1}\tau}. \tag{4}$$

In this way the vector defined by Eq. 3 is converted into a unique symbol π (see Zunino et al. (2011) and Zanin et al. (2012) for illustrative examples). In order to get a unique result we consider that $r_i < r_{i-1}$ if $x_{s-r_i\tau} < x_{s-r_i-1\tau}$. This is justified if the values of x_t have a continuous distribution so that equal values are very unusual. For all the $D!$ possible orderings (permutations) π_i of order D, their associated relative frequencies can be naturally computed by the number of times this particular order sequence is found in the time series divided by the total number of sequences. Thus, an ordinal pattern probability distribution $P = \{p(\pi_i), i = 1, \ldots, D!\}$ is obtained from the time series. This probability distribution is linked to the sequences of ranks resulting from the comparison of consecutive $(\tau = 1)$ or non-consecutive $(\tau > 1)$ points, allowing for empirical reconstruction of the underlying phase space (Bandt and Pompe 2002). It is worth noting that the method is rank-based, and the ordinal pattern probability distribution is invariant with respect to nonlinear monotonic transformations. Thus, nonlinear drifts or scaling artificially introduced by a measurement device do not modify the quantifier estimations. This property can be highly desired for the analysis of experimental data and natural time series analysis (see e.g., Carpi et al. 2010). Nevertheless, the rank-based description unavoidably results in loss of information, which is however common to every nonparametric rank-based statistical method.

The probability distribution P is obtained once we fix the embedding dimension D and the embedding delay time τ. The former parameter plays an important role for the evaluation of the appropriate probability distribution, since D determines the number of accessible states, given by $D!$ It was established that the length N of the time series must satisfy the condition $N \gg D!$ to obtain a reliable statistics (Kowalski et al. 2007; Staniek and Lehnertz 2007). As far as τ is concerned, Bandt and Pompe (2002) considered an embedding delay $\tau = 1$. Nevertheless, other values of τ might provide additional information. Soriano et al. (2011a, b) and Zunino et al. (2010a, 2012b) showed that this parameter is strongly related, when it is relevant, with the intrinsic time scales of the system under analysis.

In this work we evaluate $\mathcal{H}_S$ and $\mathcal{C}_{JS}$ using the permutation probability distribution $P = \{p(\pi_i)\colon\ i = 1, \ldots, D!\}$ so that the former quantifier is called permutation entropy and the latter permutation statistical complexity. These symbolic quantifiers were shown to be particularly useful for different purposes (Zanin et al. 2012) like characterizing stochastic processes (Rosso et al. 2007b; Zunino et al. 2008), detecting noise-induced temporal correlations in stochastic resonance phenomena (Rosso and Masoller 2009a, b), characterizing financial time series (Zunino et al. 2009, 2010b, 2011, 2012a), quantifying the randomness of chaotic pseudo-random number generators (De Micco et al. 2009), characterizing the complexity of low-frequency fluctuations in semiconductor lasers with optical feedback (Tiana-Alsina et al. 2010), and developing a formal independence test between two time series (Cánovas et al. 2011). Lange et al. (2013) used ordinal patterns and statistical complexity for analyzing a large set of worldwide river runoff time series (with a focus on the Paraná River). The present work can therefore be considered as a detailed study focused on the US rivers.

3 Complexity–entropy causality plane

3.1 CECP properties of some chaotic and stochastic signals

The CECP is the representation space obtained by plotting the permutation statistical complexity $\mathcal{C}_{JS}$ versus permutation entropy $\mathcal{H}_S$ of the system. The term causality denotes that the temporal correlation between successive samples is taken into account by using the permutation probability distribution to estimate both quantifiers.

An example is given in Fig. 1. We considered time series simulated by six types of systems, namely, the Duffing, Lorenz and Rossler chaotic systems, the Henon chaotic map, the fractional Brownian motion (fBm) stochastic process and a fully deterministic sinusoidal signal. The equations of the chaotic systems/map and their parameters' setup are detailed in the "Appendix". The sampling frequencies of the signals drawn from the chaotic and sinusoidal systems are chosen based on two criteria: (1) we used the frequencies that maximize the value of $\mathcal{C}_{JS}$

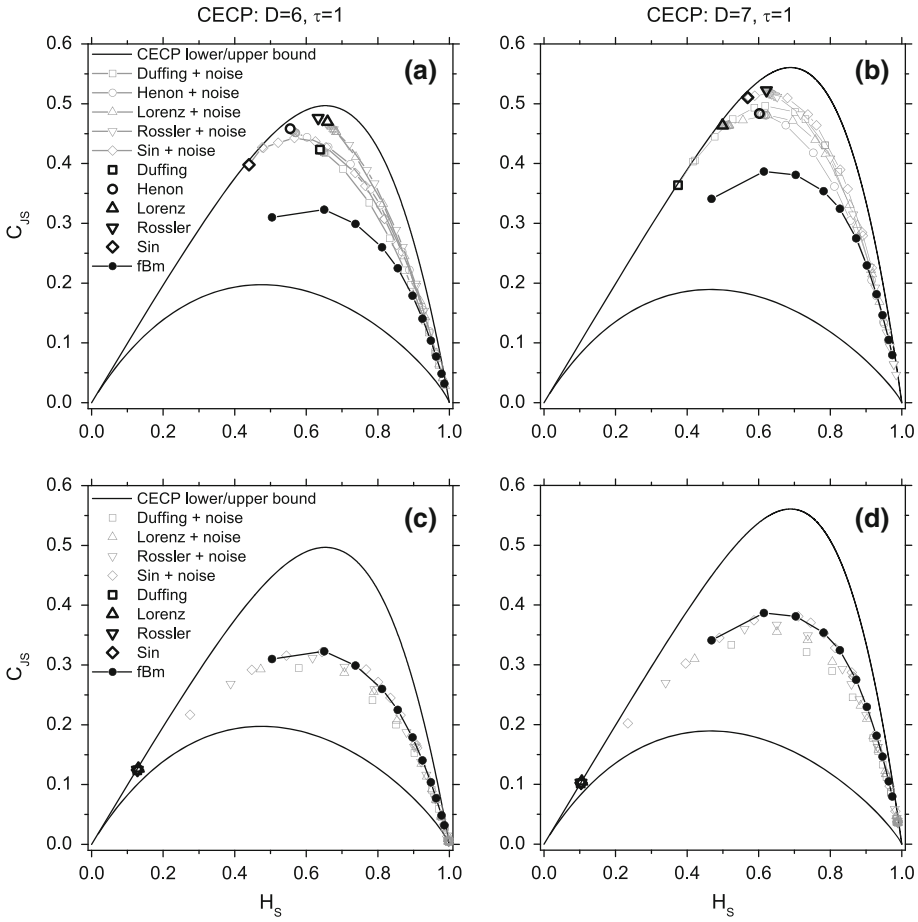

Fig. 1 CECP for signals simulated from chaotic and stochastic systems and maps. The cases $D = 6$ (*left column*) and $D = 7$ (*right column*) with $\tau = 1$ are considered. The *top panels* show the CECP patterns of the chaotic systems (Duffing, Lorenz, Rossler) optimally sampled according to the maximum achievable value of $\mathcal{C}_{JS}$ for $D = 6$, whereas the *bottom panels* show the case where the chaotic systems are sampled so that the resulting time series exhibit an average period (365 steps) and length ($\approx 75 \times 365$) similar to the seasonal components which usually characterize the observed stream flow series

for $D = 6$ and $\tau = 1$ (De Micco et al. 2010), and (2) the frequencies that yield signals which exhibit the average period or quasi-period equal to 365. The second criterion allows us to study the behavior of periodic and quasi-periodic signals that mimic the typical annual seasonality of the daily stream flow time series. In both cases, the sample size was set as 75 x 365 = 27,375, i.e. the size of the stream flow time series that are analyzed in the following sections. The sinusoidal and chaotic signals were standardized so that the amplitude $A_s = 1$. Moreover, these signals were combined with a Gaussian noise characterized by zero mean and several values of amplitude A_n ranging from 0.01 to 20, so that the signal-to-noise ratio SNR (defined as the ratio of the signal amplitude A_s to the noise amplitude A_n) ranges from 100 to 0.05. The resulting signals (sinusoidal and chaotic plus noise) allow for the assessment of the effect of the increasing level of noise on the location of the pairs $(\mathcal{H}_S, \mathcal{C}_{JS})$ in the CECP. The fBm signals were simulated using values of the power spectrum slope $\beta \in (1, 3)$, with step 0.1, corresponding to Hurst parameter values $H \in (0,1)$.

As mentioned above, for a given value of $\mathcal{H}_S, \mathcal{C}_{JS}$ can assume a range of values, which describes an admissible region in the CECP (i.e., the area enclosed between the continuous lines shown in Fig. 1). The CECP bounds only depend on D and can be computed as described by Martín et al. (2006). Focusing on the optimally sampled signals, for $\tau = 1$, Fig. 1a, b shows that the chaotic signals and the sinusoidal sequence tend to exhibit $\mathcal{H}_S < 0.7$ and the corresponding values of $\mathcal{C}_{JS}$ are close to the upper limit, meaning that the signals are characterized by mid-low entropy and high complexity (namely, the maximum complexity for the given entropy. These results agree with those obtained by Rosso et al (2007a) for other chaotic systems (logistic map, Skew Tent map and Shuster map). The additive random noise (observational noise) introduces an increasing uncertainty in the time series, reducing the signal complexity and increasing the entropy. When the noise dominates the signal, the position of the signal in the CECP tends to the boundary position $(\mathcal{H}_S = 1, \mathcal{C}_{JS} = 0)$.

Unlike the chaotic series, the fBm signals in Fig. 1a, b shows mid-high entropy, whereas the corresponding values of $\mathcal{C}_{JS}$ are in the middle between the lower and upper CECP limits. The fBm signals describe the same behavior observed for noise sequences with uniform marginal distribution and power law spectrum that were studied by

Rosso et al. (2007a, 2013). For the fBm signals, $\mathcal{H}_S$ decreases as the Hurst parameter increases, since the series exhibit an increasing persistence, which implies redundant information. For low values of H, the CECP location of the fBm approaches to the limit point (1, 0) as is expected for an uncorrelated random noise. It should be noted that these results also agree with the CECP patterns of the logistic map contaminated with additive uncorrelated noise and colored correlated noise (Rosso et al. 2012a, b).

Figure 1a, b also shows that the choice of D can change the absolute values $\mathcal{H}_S$ and $\mathcal{C}_{JS}$, but the mutual positions with respect to the CECP bounds are almost unchanged. Of course, for high values of D, the length and the number of the accessible symbols increase, leading to more accurate estimates (Zunino et al. 2011). Nevertheless, as the computational time also increases, in the analyses of observed data, we used $D = 6$ (as is also suggested by Bandt and Pompe (2002)), whereas the role of the embedding delay τ was studied by analyzing the stream flow data.

3.2 The impact of the sampling frequency on CECP

Unlike the chaotic signals, the sinusoidal signal in Fig. 1a does not exhibit a monotonic decreasing pattern as would be expected when the random noise is added. The reason of this behavior is better highlighted by the signals that are sampled with average period ≈ 365 (the second sampling strategy mentioned in the previous section). The corresponding CECP patterns (Fig. 1c, d) indicate that all signals collapse in a unique curve that characterizes the chaotic systems, the stochastic fBm and the pure deterministic sinusoidal signal. Therefore, the sampling strategy (sampling frequency) influences the capability of discrimination between chaotic and stochastic signals more than a moderate additive observational noise. However, this behavior is expected and clearly explained by De Micco et al. (2010) and Zunino et al. (2012b). The optimum sampling frequency is the minimum sampling frequency that retains all the information concerning the time structures of the signal. Using higher sampling frequencies (as it was done in the second strategy) results in an oversampling that produces unnecessary long files to cover the full attractor. In this case, specific D-length ordinal patterns $\{D-1, \ldots, 1, 0\}$ and $\{0, 1, \ldots, D-1\}$, which correspond respectively to monotonic increasing and decreasing patterns, appear more frequently than any other ordinal pattern, resulting in low entropy $\mathcal{H}_S$ and low complexity $\mathcal{C}_{JS}$. On the other hand, sampling frequencies lower than the optimal one lead to lose fundamental information concerning nonlinear correlations. In this case, since the samples are not correlated at all and the series behave randomly, all the ordinal patterns appear with almost the same frequency, resulting in high entropy $\mathcal{H}_S$ and low complexity $\mathcal{C}_{JS}$. Therefore, the points corresponding to undersampled trajectories of the continuous systems tend to fall in the region close to the limit point (1, 0). Based on these remarks, the Henon map was not reported in Fig. 1c, d, as it is a discrete system whose dynamical structure is lost for every sampling frequency lower than one.

Salas et al. (2005) already recognized the problem of the sampling frequency in the detection of the original chaotic nature of a system and showed that linear filters, such as aggregation and sampling at regular intervals, influence the values of the correlation dimension obtained by the Grassberger–Procaccia algorithm (Grassberger and Procaccia 1983) for the Lorenz system and sample size equal to 15,000. Therefore, our results generalize to some extent the finding of Salas et al. (2005), considering a different sample size, several systems and a different technique of detection.

4 Data analyses

4.1 Data

The data consist of 75 years of daily stream flow records spanning from 1934 to 2009 for 80 stations in the continental United States (Fig. 2). The dataset was retrieved from the US Geological Survey (USGS) website (http://waterdata.usgs.gov/nwis) along with the corresponding metadata. The summary statistics of the 80 series are reported in Table 1. Almost all rivers and creeks are regulated by lakes, reservoirs, power plants and diversions for irrigation, industrial and municipal supply, influencing to different extent the properties of the stream flow records. Nineteen time series show time intervals with zero stream flow values (and are denoted as intermittent). Figure 3 shows six example series along with their autocorrelation function (ACF) and power spectrum density (PSD), which describe different types of regime.

The first series refers to the Oconee River at Dublin (Georgia; USGS ID 02223500). The flow is regulated by Lake Oconee and Sinclair Reservoir and the series appears rather stationary. The ACF points out the evident annual seasonality, whereas the PSD highlights the presence of an additional meaningful weekly periodicity likely related to the lake and reservoir operation policies. Moreover, the PSD exhibits a break point at about 15-day period.

The Yellowstone River at Corwin Springs (Montana; USGS ID 06191500) is influenced by the natural storage of the Yellowstone Lake and diversions for irrigation of about 3.88 km^2, of which 0.16 km^2 are downstream of the station. The ACF and PSD highlight the dominant power of the annual component of the signal. Nevertheless, the PSD also highlights the strong inter-seasonal and intra-seasonal

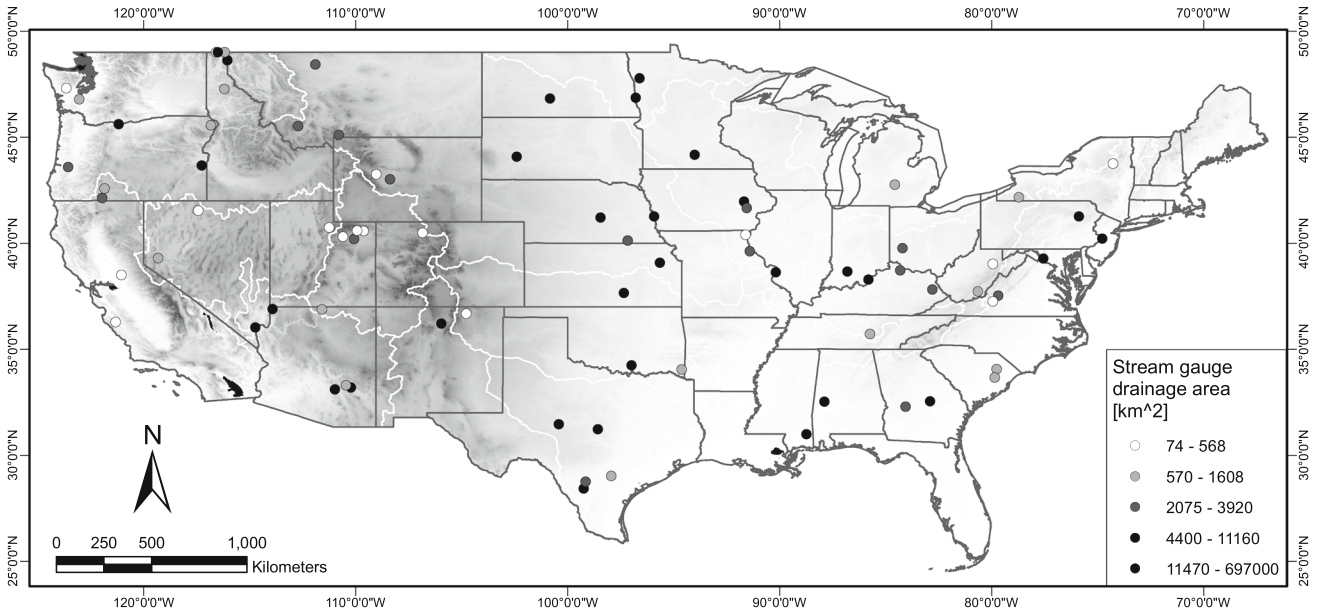

Fig. 2 Map showing the location of the USGS stream gauge stations with a record of at least 75 years used in this study. The *white lines* represent the boundaries of the first level (2-digit) hydrologic unit code

components. The time series seems not to show periodicities shorter than 60 days, which are usually related to human activities, such as power plant operation scheduling.

The Mountain Fork River near Eagletown (Oklahoma; USGS ID 07339000) exhibits a typical example of non-stationarity due to human activity, namely, the complete regulation since 1968 (day 12,410 in the figure) by the Broken Bow Lake. The seasonality is evident but the signal is more noisy than those of the Oconee and Yellowstone rivers. An apparent break point as well as a spike at 7-day period emerge in the PSD.

Nueces River at Cotulla (Texas; USGS 08194000) is an example of intermittent stream flow series showing zero stream flow values. At least the 10 % of the contributing drainage area has been regulated since March 1948. Part of the flow of the Nueces River and its headwater tributaries enter the Edwards and associated limestones in the Balcones Fault Zone that crosses the basin between Nueces River at Laguna and Nueces River below Uvalde. Losses of flow into various permeable formations occurs downstream from the Balcones Fault Zone. Many diversions for irrigation are located upstream from the station. Both ACF and PSD show that the seasonality is not so evident. The PSD exhibits an apparent break point at 7-day period. The peak at the beginning of the time series corresponds to the maximum stage recorded at this gauge station since at least 1879 ($\approx$9.87 m from floodmarks on June 18, 1935).

The flow of the Colorado River below Hoover Dam (Arizona/Nevada; USGS 09421500) is strongly affected by the intensive power plant regulations of the dam on the Lake Mead since February 1, 1935. The annual and seasonal periodicity are evident; however, they are dominated by the half-weekly and weekly periodicities, as is highlighted by the ACF and the high spikes of the PSD at 3.5- and 7-day time scales.

The Columbia River at The Dalles (Oregon; USGS 14105700) shows pattern, ACF and PSD similar to Yellowstone River; however, the considerable regulation by many large reservoirs (60 dams have a significant role in the river system management), diversions for irrigations upstream from the station, and the power plant and gates activities at The Dalles Dam (operating since March 10, 1957) cause half-weekly and weekly fluctuations (PSD spikes at 3.5- and and 7-day time scales). Moreover, the dam regulations also introduce daily fluctuations that cannot be detected in the power spectrum as their frequency is equal to the sampling frequency and the condition stated by the Nyquist-Shannon sampling theorem is not fulfilled. The behavior of the stream flow time series also exhibits a change between 1964 and 1974 (days 10,950 and 14,600 in the figure). This change is related to the flood control policies of the multiple-use reservoir storage plan developed by the US Army Corps of Engineer for the Columbia River basin after the flood of 1948 that destroyed Vanport, Oregon (Bonneville Power Administration et al. 2001). This plan has evolved over the years up to the desired level of protection of the lower Columbia. The stream flow measured at The Dalles is the control point for this operation, and, in years of low to moderate runoff, the reservoir system can be operated to limit peak flows of 12,743 $m^3 s^{-1}$ at The Dalles, the level above which damage begins to occur in areas not protected by levees. In this

Table 1 Summary statistics of the 80 USGS stream flow time series analyzed in this study

Station	USGS ID	Drainage area (km^2)	Mean BFI (−)	$x_{0.05}$ ($m^3\ s^{-1}$)	$x_{0.25}$ ($m^3\ s^{-1}$)	$x_{0.50}$ ($m^3\ s^{-1}$)	$x_{0.75}$ ($m^3\ s^{-1}$)	$x_{0.95}$ ($m^3\ s^{-1}$)	Mean ($m^3\ s^{-1}$)	Std Dev ($m^3\ s^{-1}$)
Delaware Rv @ Trenton, NJ	1315000	341.88	0.62	0.09	1.47	7.05	13.03	21.17	8.33	7.27
Susquehanna Rv @ Wilkes-Barre, PA	1463500	17,560.12	0.6	69.94	127.43	226.53	416.26	937.29	332.07	335.52
Potomac Rv @ Point Of Rocks, MD	1536500	25,796.28	0.5	35.96	90.9	202.47	467.23	1305.4	381.21	507.46
James Rv @ Buchanan, VA	1638500	24,995.98	0.52	37.94	75.32	152.34	311.48	866.49	269.32	388.77
Roanoke Rv @ Roanoke, VA	2019500	5,374.23	0.48	10.62	18.09	36.25	76.46	239.28	69.64	108.45
Oconee Rv @ Dublin, GA	2055000	1,023.05	0.52	1.7	3.11	5.86	11.5	31.71	10.57	17.98
Flint Rv @ Ga 26, near Montezuma, GA	2132000	2,667.69	0.66	5.79	10.73	19.2	36.81	84.38	29.35	30.56
Tombigbee Rv @ Demopolis L&D near Coatopa, AL	2136000	3,242.67	0.57	0.59	4.36	13.28	33.98	98.83	27.55	42.61
Pascagoula Rv @ Merrill, MS	2223500	11,395.95	0.53	21.04	44.74	82.4	161.97	419.09	137.16	169.63
Grand Rv @ Lansing, MI	2349605	7,562.77	0.56	35.54	57.06	95.14	157.87	391.06	127.93	107.02
Allegheny Rv @ Salamanca, NY	2467000	39,846.97	0.45	47.57	116.1	294.49	886.32	2576.83	677.56	886.54
Tygart Valley Rv @ Belington, WV	2479000	17,068.02	0.49	36.25	68.53	144.7	342.63	996.75	283.68	365.72
Greenbrier Rv @ Alderson, WV	3011020	4,164.7	0.46	5.97	17.13	42.76	98.83	271.27	78.5	102.42
Great Miami Rv @ Dayton, OH	3051000	1,051.54	0.31	0.62	3.88	11.47	27.24	85.8	23.3	36.14
East Fork White Rv @ Shoals, IN	3183500	3,532.74	0.37	2.86	8.92	26.62	64.85	208.41	56.63	91.8
Ohio Rv @ Louisville, KY	3212500	5,552.93	0.37	2.58	11.64	30.3	79.29	275.72	70.39	115.73
Minnesota Rv @ Mankato, MN	3253500	8,546.96	0.3	1.87	9.94	39.08	137.9	498.38	117.11	200.96
Cedar Rv @ Cedar Rapids, IA	3270500	6,503.46	0.44	7.56	14.87	30.58	66.54	256.27	65.32	104.47
Mississippi Rv @ St. Louis, MO	3294500	236,129.22	0.52	339.8	923.13	2,095.44	4,728.91	10,165.73	3,320.52	3,300.68
Red Rv of the North @ Fargo, ND	3373500	12,760.87	0.46	11.46	27.99	76.46	193.97	577.66	158.72	220.7
Red Lake Rv @ Crookston, MN	3421000	1,657.59	0.4	2.66	5.49	15.26	37.1	116.1	33.8	65.66
Marias Rv near Shelby, MT	4113000	3,185.69	0.58	4.08	8.3	15.77	31.15	73.06	24.85	27.76
Yellowstone Rv @ Corwin Springs, MT	5054000	17,611.92	0.66	0.45	3.82	9.54	19.96	67.68	19.25	37.05
Kansas Rv @ Topeka, KS	5079000	13,649.24	0.64	1.67	8.44	20.95	38.79	115.25	34.07	47.89
Arkansas Rv @ Wichita, KS	5325000	38,590.82	0.62	3.65	11.89	36.81	117.23	399.27	99.22	168.04
Rio Grande @ Embudo, NM	5454500	8,471.85	0.57	3.79	11.47	27.84	65.98	180.66	52.61	68.2
Yampa Rv @ Steamboat Springs, CO	5464500	16,860.82	0.62	15.15	31.15	61.16	124.31	348.3	107.03	140.45
Weber Rv near Oakley, UT	5495000	1,036	0.14	0.02	0.26	1.08	3.91	35.68	7.38	23.64
Columbia Rv @ The Dalles, OR	5508000	6,423.17	0.17	0.51	2.27	8.16	43.61	241.83	49.56	116.01
Umpqua Rv near Elkton, OR	6025500	6,412.81	0.73	6.06	9.34	13.34	29.17	129.12	31.41	44.43
Cosumnes Rv @ Michigan Bar, CA	6099500	8,396.74	0.7	3.45	6.57	11.33	28.32	97.98	25.94	42.39
Arroyo Seco near Soledad, CA	6191500	6,783.18	0.84	18.6	27.18	39.64	98.26	322.81	87.89	106.74
Lynches Rv @ Effingham, SC	6225000	551.67	0.68	0.42	0.91	2.92	11.61	28.6	7.78	10.53
Fox Rv @ Wayland, MO	6228000	5,980.28	0.62	2.32	7.7	11.07	19.26	94.58	22.59	33.21
Salt Rv near New London, MO	6342500	482,773.78	0.88	201.05	396.44	594.65	792.87	1,181.66	636.51	374.53

Table 1 continued

Station	USGS ID	Drainage area (km^2)	Mean BFI (−)	$x_{0.05}$ ($m^3\ s^{-1}$)	$x_{0.25}$ ($m^3\ s^{-1}$)	$x_{0.50}$ ($m^3\ s^{-1}$)	$x_{0.75}$ ($m^3\ s^{-1}$)	$x_{0.95}$ ($m^3\ s^{-1}$)	Mean ($m^3\ s^{-1}$)	Std Dev ($m^3\ s^{-1}$)
Bull Lake Ck near Lenore, WY	6423500	33,151.85	0.42	0.99	2.21	3.68	7.9	39.08	10.63	30.94
Wind Rv @ Riverton, WY	6610000	846,301.92	0.87	240.69	518.2	866.49	1,098.69	1,816.52	901.9	546.29
Missouri Rv @ Bismarck, ND	6785000	20,914.15	0.66	8.83	20.56	29.17	38.23	61.73	32.23	24.36
Cheyenne Rv near Wasta, SD	6884000	6,086.47	0.46	2.15	3.34	4.5	7.08	33.41	10.77	31.92
Missouri Rv @ Omaha, NE	6889000	146,904.13	0.58	19.82	39.64	73.62	165.37	620.14	162.46	284.45
Middle Loup Rv @ St. Paul, NE	7010000	1,805,221.71	0.83	1,682.02	2,763.72	4,304.15	6,824.35	12,827.51	5,373.88	3,617.93
Little Blue Rv near Fairbury, NE	7144300	104,868.62	0.47	2.1	5.72	12.23	25.68	116.95	29.82	63.24
Washita Rv near Dickson, OK	7203000	779.59	0.5	0.03	0.1	0.22	0.51	2.12	0.57	1.47
Mountain Fork near Eagletown, OK	7331000	18,653.09	0.37	2.44	7.73	17.1	43.32	190.29	48.31	107.58
Colorado Rv near San Saba, TX	7339000	2,038.32	0.22	0.27	4.53	13.51	39.64	165.65	38.14	76.42
Colorado Rv below Hoover Dam, AZ-NV	8136000	14,353.71	0.19	0	0.04	0.23	0.91	5.27	2.63	35.25
Gila Rv @ Kelvin, AZ	8147000	80,851.66	0.28	0.93	3.06	6.23	16.42	107.89	30.09	125.49
St Joe Rv @ Calder, ID	8186000	2,141.92	0.3	0.21	0.45	0.79	1.47	9.59	3.92	23
Concho Rv @ San Angelo, TX	8194000	13,392.83	0.15	0	0	0.03	2.32	27.82	7.36	39.62
Indian Rv near Indian Lake, NY	8205500	8,881.07	0.19	0	0	0.19	1.25	11.48	4.24	39.15
Levisa Fork @ Paintsville, KY	8279500	26,935.88	0.79	6.43	10.39	15.15	23.45	95.14	25.98	33.88
Licking Rv @ Catawba, KY	9239500	1,471.11	0.66	1.81	2.72	3.85	10.34	63.15	13.06	21.09
Collins Rv near Mcminnville, TN	9266500	261.59	0.71	0.31	0.62	1.19	2.83	11.52	2.76	4.45
Big Hole Rv near Melrose, MT	9277500	914.27	0.77	1.73	2.49	3.06	4.22	20.61	5.29	6.9
Vermejo Rv near Dawson, NM	9295000	6,845.34	0.64	0.54	2.52	7.93	12.69	52.1	13.11	21.19
Ashley Ck near Vernal, UT	9299500	282.31	0.71	0.62	0.88	1.47	3.68	12.21	3.27	4.81
Duchesne Rv near Tabiona, UT	9382000	3,651.88	0.38	0.08	0.15	0.4	0.65	2.1	0.8	3.23
Duchesne Rv @ Myton, UT	9415000	13,183.04	0.68	1.59	2.12	4.13	6.63	21.32	6.72	12.55
Whiterocks Rv near Whiterocks, UT	9421500	444,700.96	0.63	141.58	267.03	379.45	495.54	710.75	393.91	179.01
Paria Rv @ Lees Ferry, AZ	9466500	29,707.16	0.43	0	0.51	1.95	7.05	46.16	10.44	46.96
Virgin Rv @ Littlefield, AZ	9468500	2,657.33	0.31	0	0.11	0.27	0.57	4.13	1.66	11.93
Gila Rv @ Calva, AZ	9474000	46,648.28	0.6	0.17	2.61	7.22	16.83	38.51	14.52	43.04
San Carlos Rv near Peridot, AZ	10128500	419.58	0.7	1.22	1.64	2.27	4.76	28.6	6.15	9.75
Carson Rv near Fort Churchill, NV	10312000	3,372.16	0.65	0	1.02	4.81	12.4	43.89	10.82	17.28
Sprague Rv near Chiloquin, OR	10329500	453.9	0.69	0.14	0.2	0.28	0.93	4.11	0.97	1.85
Klamath Rv @ Keno, OR	11152000	631.96	0.42	0	0.17	0.79	3.09	20.33	4.76	15.86
Chehalis Rv near Grand Mound, WA	11335000	1,388.23	0.56	0.08	0.76	2.89	15.4	55.22	13.96	34.59
Wynoochee Rv above Save Ck near Aberdeen, WA	11501000	4,053.33	0.8	4.53	7.24	9.77	17.56	52.1	16.46	18.37
Kootenai Rv @ Leonia, ID	11509500	10,152.75	0.78	8.24	20.64	34.83	60.31	124.59	46.12	38.42
Moyie Rv @ Eastport, ID	12027500	2,318.04	0.46	5.13	10.58	35.11	96.56	311.48	79.67	115.94

Table 1 continued

Station	USGS ID	Drainage area (km^2)	Mean BFI (−)	$x_{0.05}$ (m^3 s^{-1})	$x_{0.25}$ (m^3 s^{-1})	$x_{0.50}$ (m^3 s^{-1})	$x_{0.75}$ (m^3 s^{-1})	$x_{0.95}$ (m^3 s^{-1})	Mean (m^3 s^{-1})	Std Dev (m^3 s^{-1})
Boundary Ck near Porthill, ID	12036000	191.92	0.56	5.01	9.46	14.89	27.84	73.34	24.51	28.44
Kootenai Rv @ Porthill, ID	12305000	30,406.46	0.78	82.12	135.35	222.57	504.04	1,330.89	398.75	421.84
Iowa Rv @ Iowa City, IA	12306500	1,476.29	0.69	1.59	2.8	5.1	18.52	92.6	19.42	31.53
Martin Ck near Paradise Valley, NV	12321500	251.23	0.68	0.54	0.93	1.7	5.15	27.01	5.7	9.22
Cibolo Ck near Falls City, TX	12322000	35,482.84	0.78	99.96	164.8	288.83	572	1,384.69	446.44	431.83
Nueces Rv @ Cotulla, TX	12414500	2,667.69	0.7	9.26	14.44	28.06	80.42	251.45	66.02	86.09
Frio Rv near Derby, TX	13183000	28,904.27	0.6	0.06	0.14	2.41	5.61	65.98	11.42	37.38
Owyhee Rv below Owyhee Dam, OR	13292000	1,610.97	0.76	2.72	4.13	6.46	18.55	50.97	14.45	16.88
Imnaha Rv @ Imnaha, OR	14105700	613,827.18	0.89	2,010.49	2,916.63	4,020.99	6,512.86	13,648.7	5,400.93	3,777.43
Black Rv @ Kingstree, SC	14321000	9,538.93	0.55	26.9	39.08	110.72	257.68	690.93	208.74	303.9

The symbol x_P, with $P \in \{0.05, 0.25, 0.50, 0.75, 0.95\}$, denotes the quantile with nonexceedance probability P, while "Std Dev" is the standard deviation

context, large dams were built until the middle of the 1970s.

4.2 Dependence of $\mathcal{H}_S$ and $\mathcal{C}_{JS}$ on τ and the effect of the deseasonalization

In this section, we study the dependence of $\mathcal{H}_S$ and $\mathcal{C}_{JS}$ on the embedding delay time τ used to obtain the ordinal pattern probability distribution defined in Sect. 2.2 for $D = 6$. The analysis is carried out on the observed stream flow series and the corresponding deseasonalized sequences. The deseasonalization was performed by a three-step procedure. Firstly, the average discharge and the standard deviation for each calendar day were computed. Secondly, the obtained seasonal patterns were smoothed applying the local weighted scatterplot smoothing (LOESS) filter (Cleveland and Devlin 1988). Finally, each observation was standardized through the relationship $Y_i(t) = [X_i(t) - m_i] / s_i$, for $i = 1, \ldots, 365$, where m_i and s_i are the values of the smoothed average and standard deviation for a given calendar day. We notice that the deseasonalized signals preserve the other multiple periodic trends that are often shown by the stream flow time series at different time scales. Nevertheless, as the annual periodicity is a common property of daily stream flow records and the deseasonalization is a common preprocessing procedure (e.g., Montanari et al. 1997, 2000; Kantelhardt et al. 2003; Montanari 2005; Koscielny-Bunde et al. 2006) it is worth studying its effect on the complexity and entropy quantifiers.

Figure 4 shows the patterns of $\mathcal{H}_S$ and $\mathcal{C}_{JS}$, for $\tau = 1, \ldots, 800$, corresponding to the six series described in the previous section for raw and deseasonalized data. For the Oconee River and Yellowstone river, the patterns of $\mathcal{H}_S$ and $\mathcal{C}_{JS}$ tend to show well defined peaks corresponding to values of τ multiple of 365 (the annual cycle), meaning that the signal tends to be less organized as the delay between the observations matches multiples of 1 year. Even though this result might be expected, we note that these peaks emerge from an underlying plateau and their magnitude reflect the strength of the annual component of the signal in terms of information content. Indeed, after deseasonalization, the plateau shifts towards the peak values, meaning that the removal of the annual periodicity returns a more random signal with similar behavior for all τ values.

For the Oconee River, the contribution of the seasonality is less relevant than in the Yellowstone River, as already highlighted in Fig. 3. The seasonality exerts even less influence on the temporal structure of the Mountain Fork River. We notice that the magnitude of the peaks follows the same pattern of the amplitude of the ACF waves in Fig. 3.

The $\mathcal{H}_S$ and $\mathcal{C}_{JS}$ patterns of the Nueces River signal do not exhibit peaks at any scale, meaning that there are not dominant time scales, as is also pointed out by the PSD.

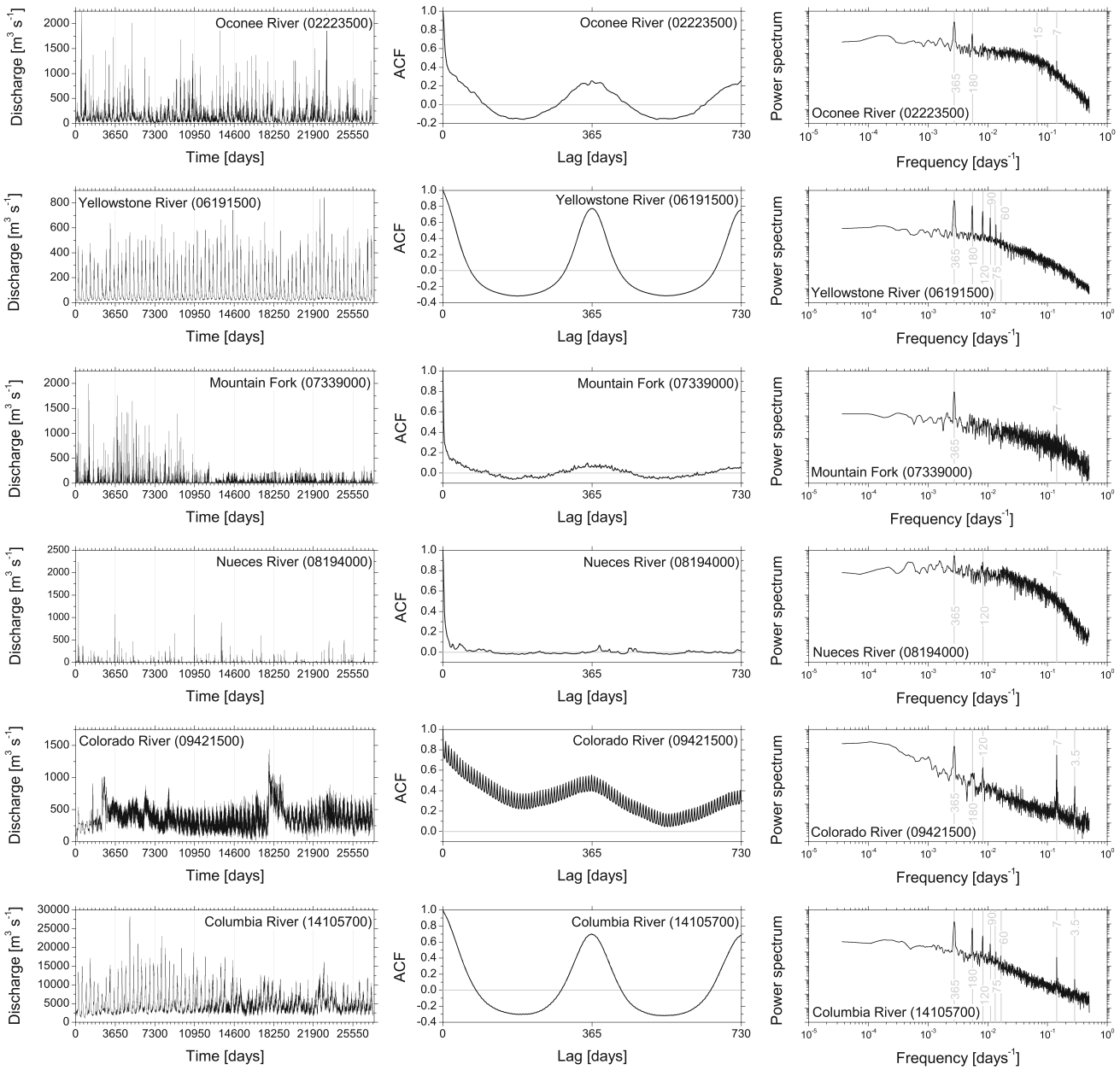

Fig. 3 Times series (*panels on the left*), ACF (*panels in the middle*) and PSD (*panels on the right*) of the six example stream flow series that are described in detail in the text

The contribution of the seasonal component is highlighted by the magnitude of the shift of the patterns after deseasonalization. In this case, the peaks characterize the patterns of the deseasonalized series. This behavior is related to the intermittent nature of this time series and is further studied in Sect. 4.5.

The $\mathcal{H}_S$ and $\mathcal{C}_{JS}$ patterns of the Colorado River are characterized by evident weekly oscillations (multiple of $\tau = 7$). The small shift after deseasonalization and the high (small) values of $\mathcal{H}_S$ ($\mathcal{C}_{JS}$) indicates that the signal is dominated by weekly periodicity. However, the weekly regulation is not able to introduce a strong temporal organization in the signal, which shows a pattern characterized by changes that are related to particular dam operation activities (Fig. 3 bottom-left panel). The Columbia River exhibits $\mathcal{H}_S$ and $\mathcal{C}_{JS}$ patterns similar to Oconee River with weekly oscillations superimposed to the evident seasonality.

It is worth noting that the values of $\mathcal{H}_S$ and $\mathcal{C}_{JS}$ of the raw and deseasonalized series are identical (apart from numerical approximations) for multiples of $\tau = 365$, as the sequences $\{Y_t, Y_{t+365}\ldots, Y_{t+365D}\}$ in the reconstructed phase space are identical to $\{X_t, X_{t+365}\ldots, X_{t+365D}\}$ unless a normalizing factor that is the same for all values of the set

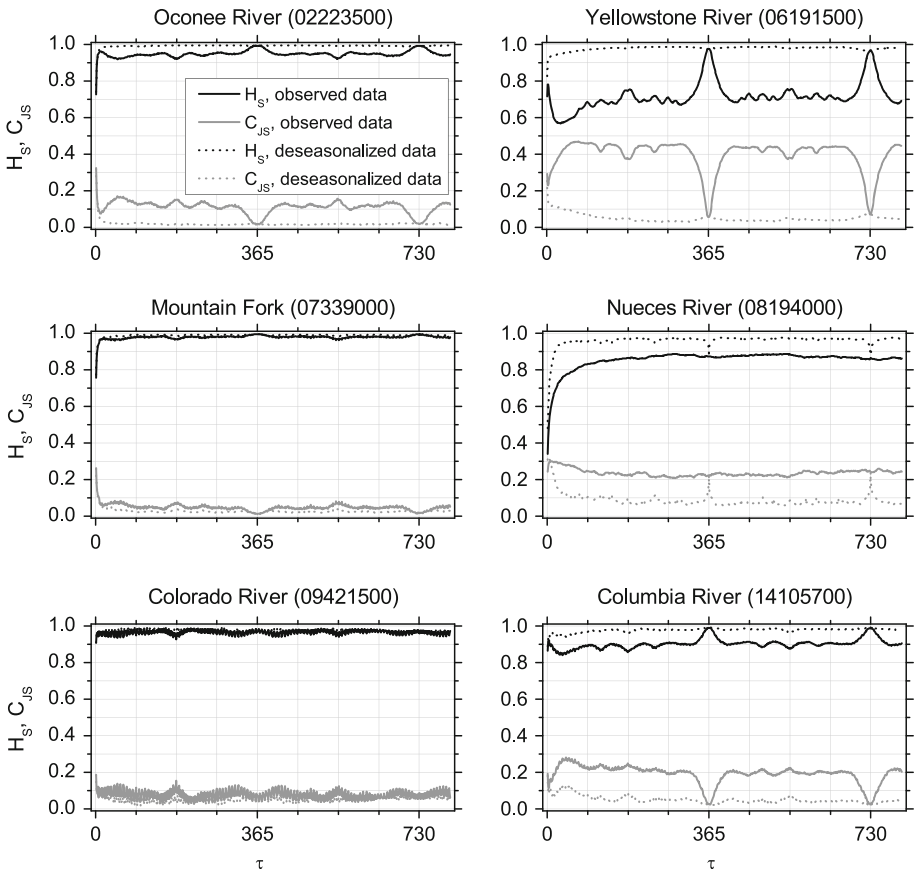

Fig. 4 Patterns of $\mathcal{H}_S$ and $\mathcal{C}_{JS}$ for $D = 6$ and $\tau = 1, \ldots, 800$. The figure refers to the six example stream flow series shown in Fig. 3

(Fig. 5). Moreover, $\mathcal{H}_S$ and $\mathcal{C}_{JS}$ values for $\tau = 1$ are also weakly influenced by deseasonalization because the values of the sequences $\{Y_t, Y_{t+1}\ldots, Y_{t+D}\}$ are computed using the values of the seasonal components of D consecutive days, which in turn are reasonably close to each other (Fig. 5).

4.3 CECP analysis

The CECP analysis is based on the values of the complexity quantifiers for $D = 6$ and two values of the embedding delay τ, namely, 1 and 150. The first value of τ allows us to recognize the short-term structure of the signals, whereas the values of $\mathcal{H}_S$ and $\mathcal{C}_{JS}$ for $\tau = 150$ represent the overall structure of the series for several values of τ, as is testified by the approximate plateaus emerging in Fig. 4. The results for $\tau = 1$ are displayed in Fig. 6, where the theoretical patterns of Fig. 1c are shown in gray for the sake of comparison. The position of the 80 observed series in the CECP (Fig. 6a) indicates that the structure of the stream flow series is close to that of the fBm (with Hurst parameter $H > 0.5$) and the noisy oversampled sinusoidal and chaotic sequences. This agreement confirms the difficulty in characterizing real world data in terms of chaotic or stochastic processes. On

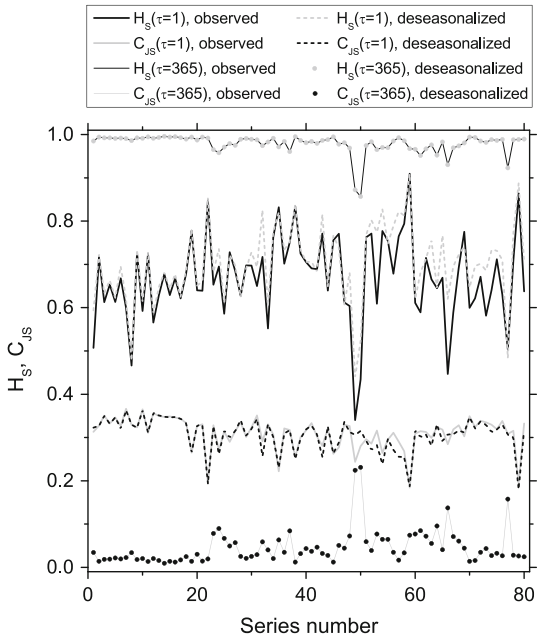

Fig. 5 $\mathcal{H}_S$ and $\mathcal{C}_{JS}$ values of observed and deseasonalized stream flow series for $D = 6$ and $\tau = \{1,365\}$

the other hand, the "middle" position of the points denotes that the distinction between chaotic or stochastic interpretation may be ill-posed because of the interplay of

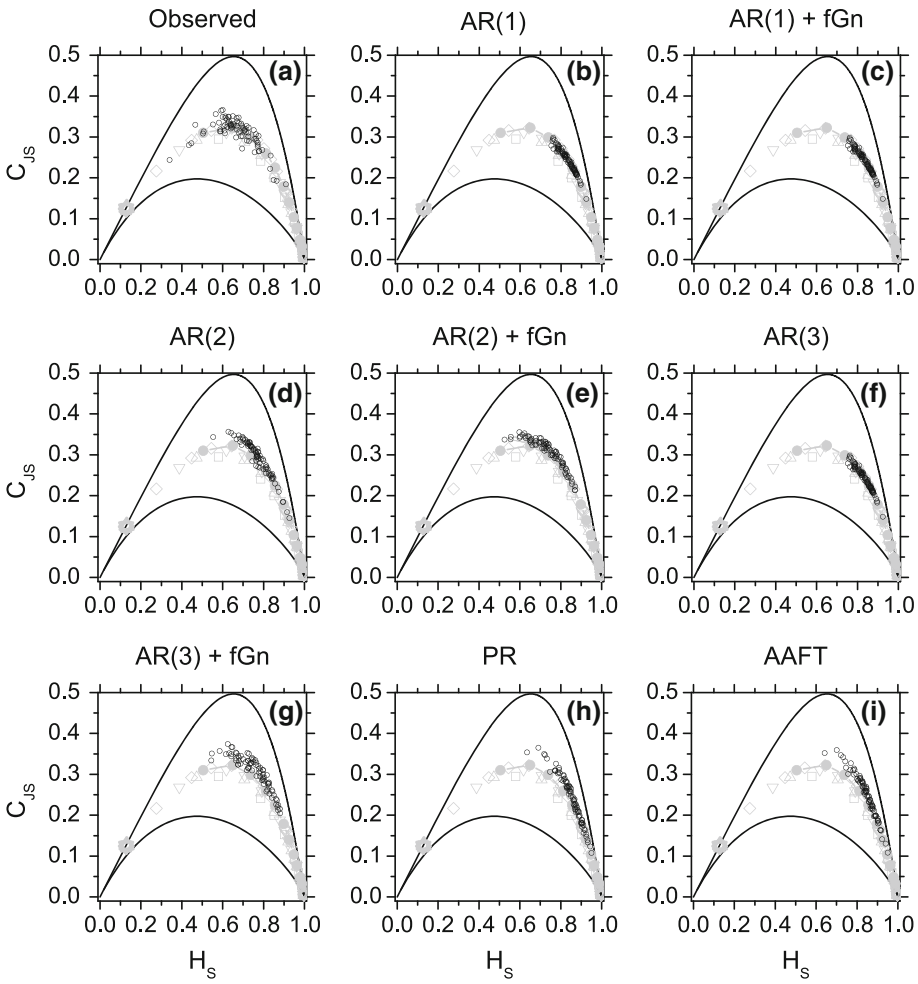

Fig. 6 CECP patterns ($D = 6$, $\tau = 1$) of the observed stream flow time series and the corresponding surrogate sequences simulated by eight different methods. *Gray lines* and *symbols* denote the patterns corresponding to the theoretical models shown in Fig. 1c. CECP patterns for the surrogate data correspond to the average values over 100 simulations. Sampling error is not shown as it is negligible

multiple factors, such as sample size, intermittency and sampling frequency. These findings can partially explain the variety of results on the recognition of chaotic/stochastic behavior in some geophysical time series: if the series fall in "shared" zone of the CECP, both mechanisms (low- and high-dimensional) can be plausible under some conditions and can be reasonably detected. The above results further stress the importance of the sampling procedure recognized by Salas et al. (2005), which therefore should be added to the list of the key factors that influence the analysis of the signal dynamics.

To better understand the temporal structure of the stream flows sequences, we have computed the values of $\mathcal{H}_S$ and $\mathcal{C}_{JS}$ of synthetic series simulated by different methods that are commonly applied to emulate stream flow series. We have considered linear auto regressive (AR) models of order 1–3, the Theiler phase randomization (PR) and the Theiler amplitude adjusted Fourier transform (AAFT; Theiler et al. (1992)). PR and AAFT are two nonparametric techniques that are widely used to generate surrogate series which preserve the observed power spectrum by resampling data in the frequency space. The AR synthetic series were simulated using two types of innovation values: the resampled residuals obtained during the model calibration, and innovations drawn from a fractional Gaussian noise (fGn, also known as Hurst–Kolmogorov process (e.g., Koutsoyiannis 2011a; Tyralis and Koutsoyiannis 2011)). The second method results in the AR-fGn models suggested by Király and Jánosi (1998) and similar to that proposed by Livina et al. (2003). When long range dependence is detected in the time series (as it is discussed in the following sections), these models allow us to introduce it in the simulated time series in a simple manner. As the CECP is a rank-based technique, we do not aim at faithfully reproducing the observed series, but the overall temporal structure of the signals.

For $\tau = 1$ (Fig. 6b–i), the AR1, AR1-fGn, AR2 and AR3 models seem to be too simple, whereas the AR2-fGn and AR3-fGn signals exhibit a degree of complexity closer to the observed series. It is worth noting that reproducing the PSD (PR method) and the amplitude distribution and PSD (AAFT method) is not sufficient to preserve all the information embedded in the original series, which are characterized by deterministic changes and possible periods of zero stream flow. These properties introduce additional temporal structures that

are not modeled by the simple models mentioned above. The use of fGn innovations results in an apparent improvement of the CECP patterns. Moreover, the AR-fGn models do not account for intra-annual periodicities, which can play an important role in the overall structure of the signals. Nevertheless, the results show that the short-term behavior of daily stream flows series is close to that of some simple stochastic time series. The CECP patterns corresponding to $\tau = 150$ (Fig. 7) highlight that all methods but AR2-fGn and AR3-fGn reproduce the CECP patterns. In this case, setting $\tau = 150$ smoothed out the signal components with frequency <150 days (e.g., weekly cycles), thus making the reproduction of the temporal patterns corresponding to $\tau = 150$ easier than the simulation of the pattern related to $\tau = 1$.

To further highlight the role of the seasonality on the temporal structure of the daily stream flow signals, Fig. 8 shows the CECP patterns of the observed series, their quasi-periodic LOESS smoothed mean component, and the deseasonalized series. For $\tau = 1$, the points of the seasonal components are close to the CECP upper bound and the position of the pure oversampled sinusoidal/chaotic signals shown in Fig. 1c, whereas the points corresponding to the deseasonalized series are close to those of the observed series, as already shown in Fig. 5. For $\tau = 150$, the points of the seasonal components are still close to the CECP upper bound (as expected); however, in this case, the mutual positions of the points corresponding to the three series (observed, seasonal components and deseasonalized) highlight the strong impact of the seasonality on the signal structure. Recalling that the values of $\mathcal{H}_S$ and $\mathcal{C}_{JS}$ for $\tau = 150$ are a good approximation of the plateau values that characterize a signal (Fig. 4), these patterns also hold for other values of τ. Therefore, we can argue that the deseasonalized daily series reasonably behave similar to stochastic processes as they fall in the CECP region close to limit point (1, 0). Nevertheless, as this result relies on a graphical analysis, we cannot exclude that the position of the deseasonalized series results from a chaotic process affected by a dominant noise with small SNR. However, this does not change our conclusions because in this case the signals are more similar to those corresponding to stochastic rather than deterministic processes.

4.4 Empirical relationships between CECP quantifiers, Hurst parameter H, and other stream flow properties

The Hurst parameter is a well-known index that measures the long range dependence of a process (e.g., Beran 1994), and characterizes important theoretical stochastic processes, such as fGn and the corresponding fBm. Bandt and Shiha (2007) showed that some analytical relationships do exist between H and the probability distribution $P = \{p(\pi_i),\ i = 1, \ldots, D!\}$, and therefore $\mathcal{H}_S$ and $\mathcal{C}_{JS}$, for fGn and fBm processes, and derived explicit formulas for $D = \{3,4\}$ and arbitrary $\tau > 0$ (see also Zunino et al. 2008). In this study, we investigate the empirical relationship among the CECP quantifiers, the Hurst parameter H and other physical attributes of the stream flow series on a real world dataset.

4.4.1 *Hurst parameter*

Following Serinaldi (2010), the Hurst parameter H was computed through three different methods based on different rationales to check the reliability of the estimates. We used the aggregate variance method (AGV; Beran (1989)), the detrending fluctuation analysis (DFA; Peng et al. (1994)) with a 3-order smoothing polynomial and Higuchi method (HIG; Higuchi (1988)). The values of H were computed on the deseasonalized series to avoid the detrimental effects of the seasonality (Hu et al. 2001; Marković and Koch 2005; Ludescher et al. 2011; Zhang et al. 2011). The results are shown in the scatter plot matrix of Fig. 9. The three methods provide coherent results (see the upper triangular matrix of plots). All methods yield values >0.5. The range of H values is consistent with the results published in the literature for stream flow series from international rivers (Kantelhardt et al. 2006; Wang et al. 2007; Movahed and Hermanis, 2008; Hirpa et al. 2010; Rego et al. 2013). The HIG method yields H values which are slightly different from those obtained by the AGV and DFA methods. Distinguishing perennial stream flow series (empty circles) and intermittent series (filled circles), we note that the discrepancy mainly refers to the H estimates corresponding to the intermittent sequences. As the HIG method is based on the computation of the fractal dimension, while the others rely on the variance of the mean (AGV) and the mean of the variance (DFA), the algorithm is probably sensitive to strings of zero values, which contribute to the persistence of the signal, inflating the values of H. However, we do not further explore this aspect and assume the DFA values for the following analysis because this method provides rather reliable results for stationary and nonstationary time series (e.g., Kantelhardt et al. 2002; e.g., Serinaldi 2010).

4.4.2 *Relationships between CECP and stream flow attributes*

The scatter plot matrix in Fig. 10 shows the pairwise relationships among seven stream flow attributes. We report results corresponding to $\mathcal{H}_S$ and $\tau = 150$ for the deseasonalized series since dual conclusions hold for $\mathcal{C}_{JS}$.

A negative relationship does exist between $\mathcal{H}_S$ and H, except for an isolated point referring to the Owyhee River below the Owyhee Dam. This time series is strongly

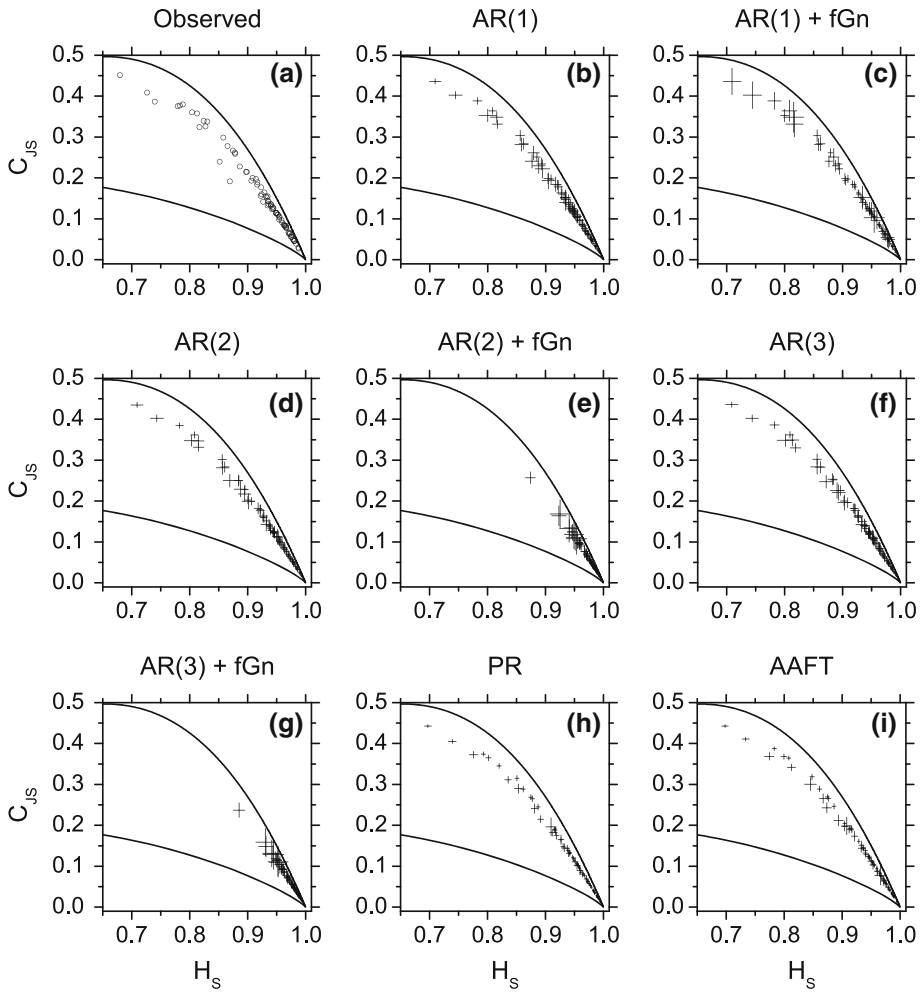

Fig. 7 CECP patterns ($D = 6$, $\tau = 150$) of the observed stream flow time series and the corresponding surrogate sequences simulated by eight different methods. CECP patterns for the surrogate data correspond to the average values over 100 simulations. The length of the segments describes the sampling standard error around the average

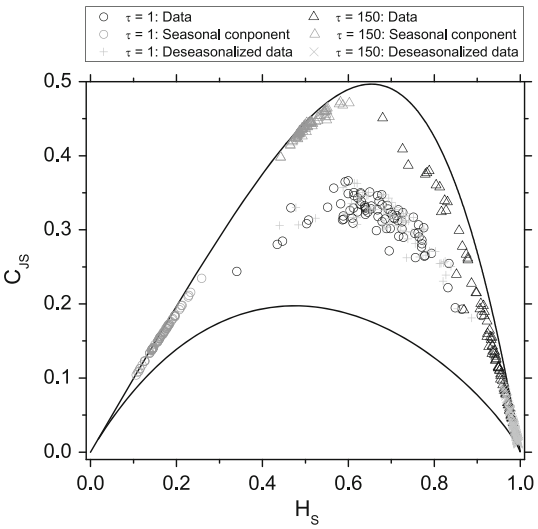

Fig. 8 CECP patterns of the 80 observed stream flow time series (*black symbols*) and the corresponding mean seasonal components (*gray symbols*) and deseasonalized sequences (*light gray symbols*) for $D = 6$ and $\tau = \{1,150\}$

influenced from the reservoir operations scheduling, which results in a well-defined short-term temporal structure and a small $\mathcal{H}_S$ value. Distinguishing perennial and intermittent stream flow series, the LOESS curve computed on the perennial points shows that the relationship is monotonic. The values of the corresponding Kendall correlation coefficient, which are shown in the upper triangular matrix of plots, are significant at 5 % level of significance, and confirm the strength of the relationship between $\mathcal{H}_S$ and H. Therefore, more persistent series generally exhibit lower entropy and higher complexity (figure not shown). The intermittent series tend to exhibit low H and low $\mathcal{H}_S$ (high $\mathcal{C}_{JS}$) as the zeros introduce a set of preferential accessible states.

A non-monotonic relationship characterizes $\mathcal{H}_S$ and the logarithm of the drainage area $\log(A)$. In this case, the Kendall correlation coefficient is statistically significant at the 5 % level and negative (τ_K is positive for $\mathcal{C}_{JS}$ and $\log(A)$). Therefore, we may argue that the degree of randomness of the stream flow signal resulting from small drainage areas is larger than that corresponding to large drainage areas.

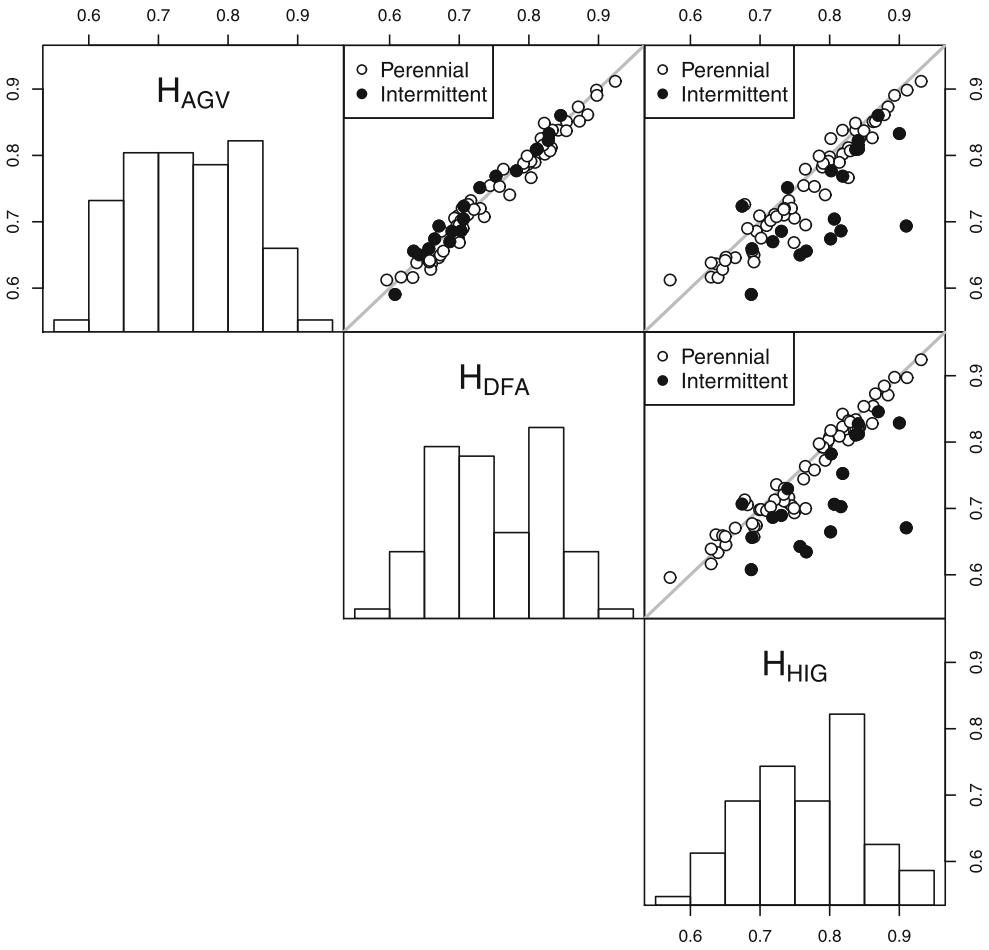

Fig. 9 Hurst parameter values of deseasonalized series computed by aggregate variance (AGV) method, detrending fluctuation analysis (DFA) with 3-order polynomial and Higuchi (HIG) method. Perennial and intermittent stream flow series are highlighted by *empty* and *filled circles*, respectively

However, the non-monotonic pattern of the relationship between $\mathcal{H}_S$ and $\log(A)$ does not allow us to confirm this statement, which must be considered with care.

For the logarithm of median, 95th percentile, and standard deviation, the patterns of the LOESS curves of the pairs $(\mathcal{H}_S, \log(x_{0.5}))$, $(\mathcal{H}_S, \log(x_{0.95}))$ and $(\mathcal{H}_S, \log(\mathrm{SD}))$ are similar to that of the pairs $(\mathcal{H}_S, \log(A))$, confirming the presence of non-monotonic relationships. Moreover, for these summary statistics, the Kendall correlation is significant at the 5 % level and positive (negative for $\mathcal{C}_{JS}$). However, the non-monotonic patterns make the values of the Kendall correlation questionable. Therefore, our results must be considered with care and should be only used as a base for more extensive analyses.

4.4.3 *Relationships between stream flow attributes with emphasis on the association between H, drainage area A and BFI*

In general, in a stream flow signal (hydrograph) two components can be recognized: the base flow, which is mainly related to slowly varying ground water processes, and the quick flow (or direct runoff), which represents the direct catchment response to rainfall events (Smakhtin 2001). As the BFI denotes the percentage of runoff corresponding to the base flow, the stream flow of the rivers characterized by high BFI is likely driven by the pattern of the base flow rather than by the direct runoff. The resulting signals are therefore dominated by the components that evolve at the time scales which are typical of ground water fluxes. Thus, the negative (positive) and significant relationship between $\mathcal{H}_S$ ($\mathcal{C}_{JS}$) and BFI (Fig. 10) reflects the link between the degrees of randomness and complexity (measured by the CECP quantifiers) and the time scales of evolution of base flow and direct flow. The intermittent rivers are an exception. It is worth noting that the $(\mathcal{H}_S, \mathrm{BFI})$ panel allows us to discriminate between intermittent and perennial stream flow (in terms of separate clusters) more clearly than the other pairwise scatter plots.

The scatter plots in Fig. 10 highlight other interesting properties of the daily stream flow series. The weak relationship between H and the drainage area is coherent with the results discussed by Koscielny-Bunde et al. (2006) and Wang et al. (2007) and can be likely ascribed to the inter-basin variability, which was not taken into account in these

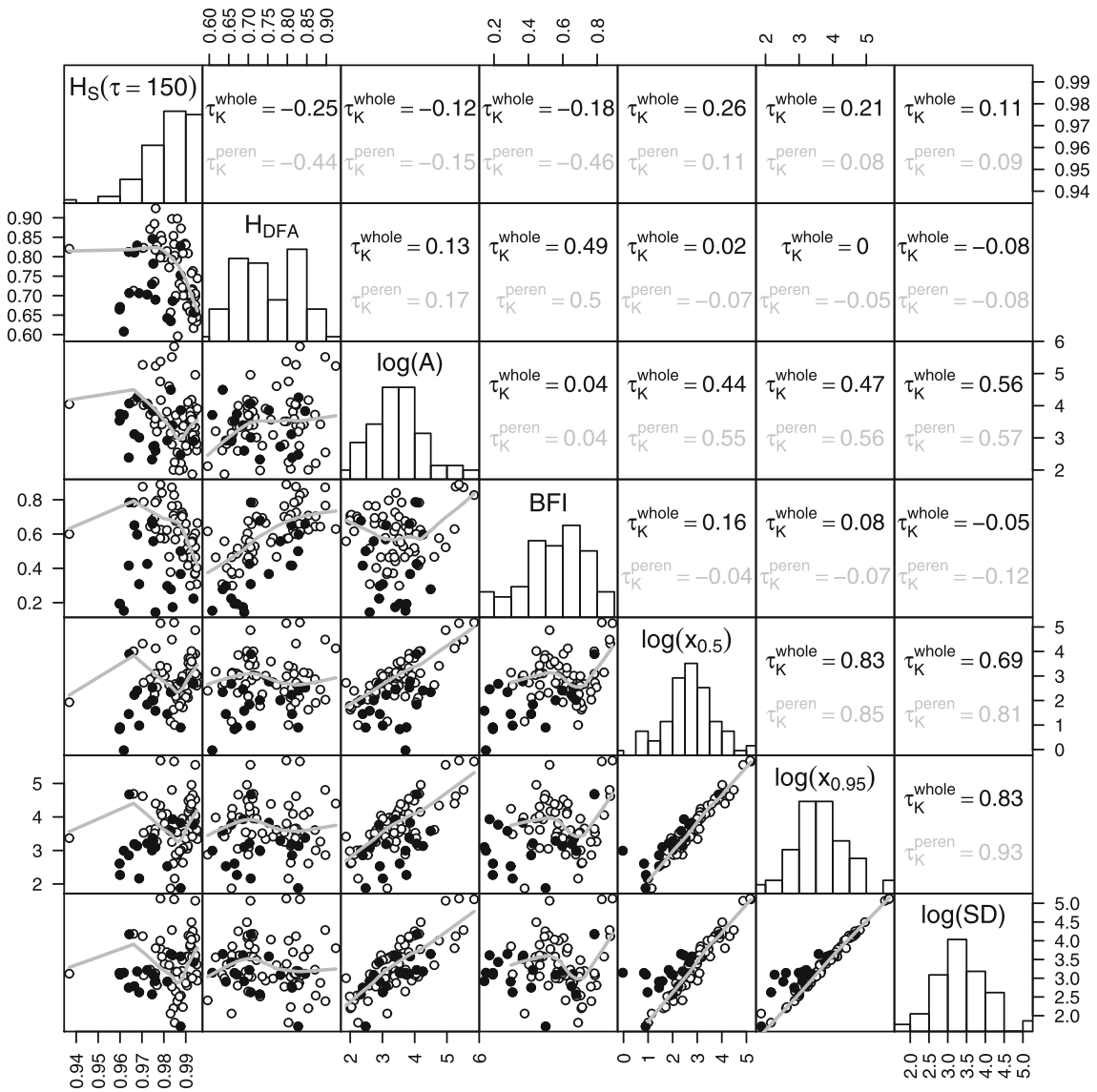

Fig. 10 The *lower triangular matrix* of panels shows the pairwise scatter plots of seven properties of the 80 observed stream flow series, namely: the $\mathcal{H}_S$ corresponding to $D = 6$ and $\tau = 150$ (D_{CECP}), the Hurst parameter computed by DFA-3 (H_{DFA}), the logarithm of the drainage area of the basin upstream each station [log(A)], the base flow index (BFI), the logarithm of the median, 95th percentile, and standard deviation of each series [log($x_{0.5}$), log($x_{0.95}$) and log(SD), respectively]. Perennial stream flow series (*empty circles*) are distinguished from the intermittent series (*filled circles*). The LOESS curves fitted to perennial stream flow series are also shown. The main diagonal shows the histograms of each properties, while the *upper triangular matrix* of panels displays the values of the pairwise Kendall correlation coefficient computed on all the points ("whole") and perennial stream flow series ("peren")

studies (Mudelsee 2007). Indeed, Mudelsee (2007) and Hirpa et al. (2010) showed that H increases with the drainange area, moving downstream along the same river basin. In particular, Mudelsee (2007) explained this behavior as the result of the aggregation of short-memory contributions from an increasing number of hydrological units.

Another quantity that in principle might increase by moving downstream is the base flow. However, while the area summarizes the filter (the basin) of the precipitation input, the base flow can be seen as a footprint of the output, namely, of the slowly varying long-term basin response. The positive and significant relationship between H and BFI values confirms that the base flow can be considered as a proxy attribute which is less affected by the inter-basin variability, as it can be related to the long-range behavior more closely than the drainage area. As a proof of concept, Fig. 11 (top panel) shows the H values reported by Hirpa et al. (2010) for 14 stream gauge stations across the Flint River Basin (Georgia, south eastern US) along with the corresponding BFI versus the drainage area. Despite the zero BFI values of the smallest sub-basins, the pattern of

the BFI tends to follow that of H. The bottom panel of Fig. 11 shows the relationship between H and BFI, and the values of Kendall correlation coefficient τ_K confirm that these attributes are related more closely than the pairs (H, A) and (BFI, A). The τ_K value was computed on the pairs (BFI, H) with and without the points corresponding to BFI = 0, to assess their impact on the estimates. It should be noted that the τ_K value for the pair (H, A) is slightly greater than that reported by Hirpa et al. (2010), probably because of small differences in the H values that can influence the ranks of the observations and the final τ_K value in this small sample.

The basin area is strongly related to the stream flow quantiles ($x_{0.50}$ and $x_{0.95}$) and the SD, which summarize the overall hydrograph (base flow plus direct runoff), while H and BFI are only weakly related to these statistics, showing no clear patterns. Therefore, the drainage area can be considered as a suitable index of the magnitude of the overall short-term response of the basin, while BFI is more related to long range dependence as it summarizes slowly-varying low-frequency ground water processes.

Finally, we notice that the SD values of the intermittent rivers tend to be larger than that of the perennial streams for similar values of $x_{0.50}$ and $x_{0.95}$. This behavior depends on zero values: if a perennial series and an intermittent series show the same median, therefore the latter must show higher variability to compensate the null contribution of zero flow records.

4.5 Comparison of simple linear and nonlinear diagnostics: the example of the Nueces River at Cotulla

The results discussed in the previous sections highlight the importance of using several techniques to study the complex behavior of daily stream flow series. In this section, we further stress this point showing the different information which can be provided by a few linear and nonlinear diagnostic plots applied to the deseasonalized series of the Nueces River at Cotulla. Figure 12 shows the ACF, the average mutual information (AMI), the PSD and the so-called lag plot. ACF and PSD are linear measures; they convey the same information, but PSD can highlight the existence of periodic components at middle and high frequencies better than ACF (see e.g. Fig. 3). The AMI for measuring the lag-1 temporal dependence of discrete time series is computed as:

$$\begin{aligned}\mathcal{I}(x_t, x_{t+k}) &= \sum\sum p_2(x_t, x_{t+k}) \log \frac{p_2(x_t, x_{t+k})}{p(x_t)p(x_{t+k})} \\ &= S(x_t) - S(x_t|x_{t+k}) \\ &= S(x_t) + S(x_{t+k}) - S(x_t, x_{t+k}),\end{aligned} \quad (5)$$

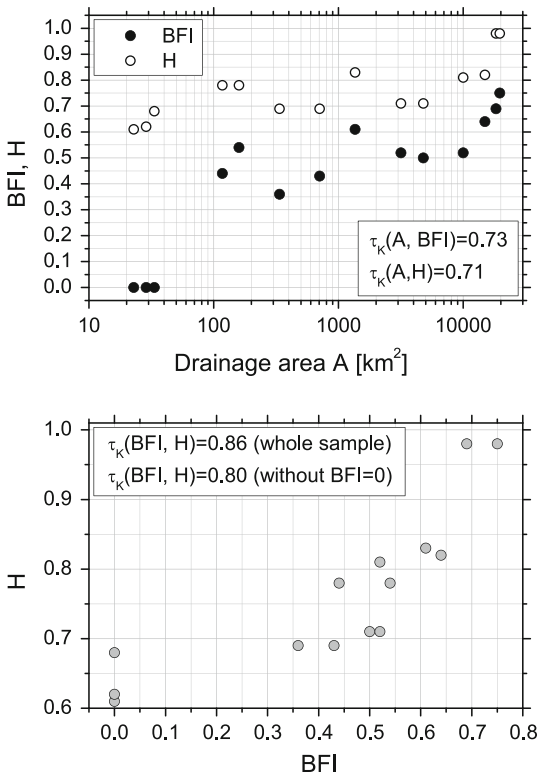

Fig. 11 *Top panel* scatter plots of BFI and H versus drainage area A for the Flint River basin analyzed by Hirpa et al. (2010). *Bottom panel* scatter plots of BFI versus H. Values of Kendall rank correlation coefficient τ_K are also shown. The τ_K values for the pair (BFI, H) are computed for the whole sample and excluding the points characterized by BFI = 0

where p_2 is the joint distribution of the the variables X_t and X_{t+k}, k is the lag, $S(x_t)$ and $S(x_{t+k})$ are the marginal entropies, $S(x_t|x_{t+k})$ is the conditional entropy and $S(x_t, x_{t+k})$ denotes the joint entropy. The AMI is a nonlinear measure like $\mathcal{H}_S$ and $\mathcal{C}_{JS}$. The lag plot is the plot of the pairs (x_t, x_{t+k}), it enables the detection of nonlinearities and was suggested by Sivakumar et al. (2007) and Sivakumar and Singh (2012) as a possible tool for classifying stream flow time series.

Based on Fig. 12a, the ACF of the deseasonalized stream flow series exhibits strong persistence; a residual seasonality can exist but is difficult to recognize. The PSD (Fig. 12b) is approximately linear, and the peak close to the annual frequency ($\approx 2.74 \times 10^{-3}$) is not prominent. On the other hand, the AMI plot (Fig. 12c) highlights the presence of a strong annual periodicity superimposed to higher frequency cycles. The pattern of the AMI is very similar to the patterns of $\mathcal{H}_S$ and $\mathcal{C}_{JS}$ shown in Fig. 4 since these quantifiers are devised to recognize the temporal structure of the time series. The lag plot (Fig. 12d) indicates that the process is the sum of two processes: the first one is the quasi-periodic seasonal component resulting

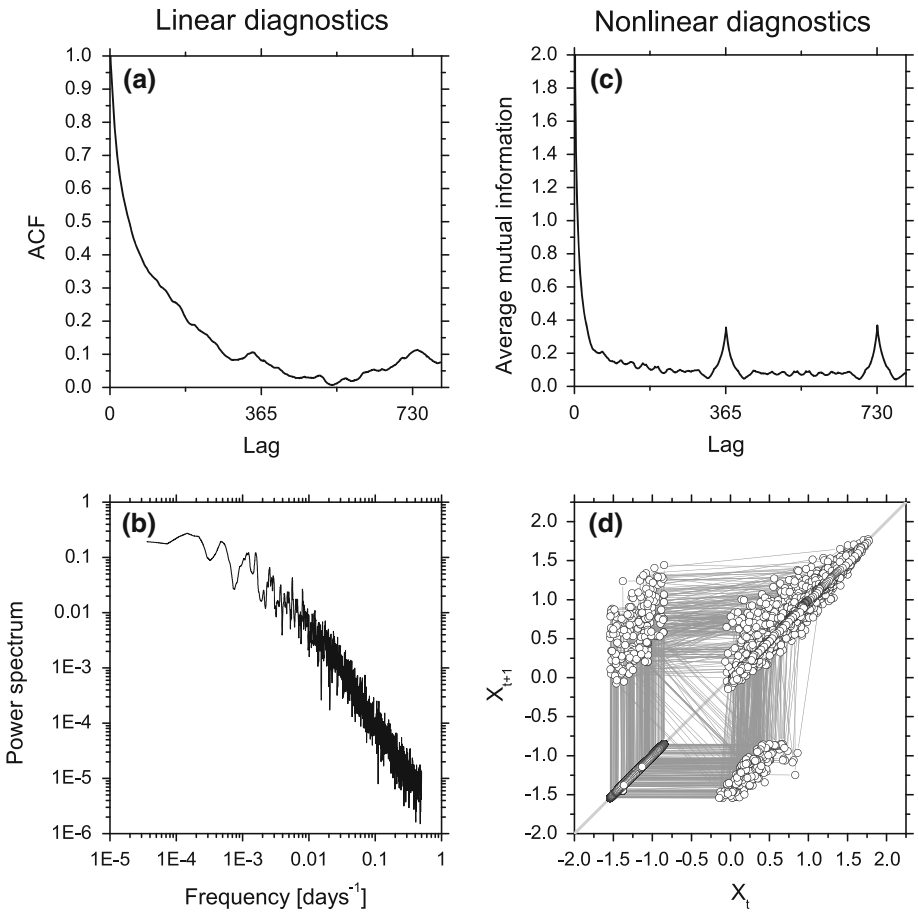

Fig. 12 Linear and nonlinear diagnostics for the Nueces River deseasonalized data: **a** ACF, **b** PSD, **c** AMI, **d** lag plot

from the deseasonalization of the sequences of zero flow, while the second one represents the deseasonalized positive stream flow values. The two processes are shown in Fig. 13. The quasi-periodic signal corresponding to the observed zero flow is shown in black in the top panel and was selected by properly thresholding the time series. This process is magnified in the middle panel and its nature appears more clearly in the bottom panel after removing the time intervals characterized by the superposed process.

Even though the residual seasonality in the deseasonalized time series is an artifact due to the deseasonalization of the sequences of zero flow, the nonlinear techniques are able to recognize its presence, whereas linear diagnostics appear to be insensitive because they are based on the signed values of the time series. When the amplitude of the signal at given frequencies is small, as is in the case of the transformed zero values, the linear measures are not able to recognize those temporal structures. On the other hand, the measures based on probability and ranks are able to identify subtle (small amplitude) regular patterns within the signal. These results confirm the importance of using a suitable set of linear and nonlinear diagnostics for the preliminary characterization and classification of hydrological signals (Sivakumar et al. 2007).

5 Discussion

Based on the study of the synthetic series drawn from stochastic and chaotic deterministic systems, we found that the CECP can distinguish the two types of signals based on the position of the complexity quantifiers on the plane. A moderate additive Gaussian noise does not affect the results very much. On the other hand, an appropriate sampling frequency plays a fundamental role. Under oversampling conditions, the chaotic signals cannot be distinguished from purely deterministic signals. An important consequence of this behavior is that a stream flow series can be interpreted as either the composition of quasi-deterministic seasonal components and stochastic noise (with or without memory) or noisy chaotic systems. Even though, at the first glance, this lack of discrimination can appear as a shortcoming of the CECP, actually, it confirms the findings of Regonda et al. (2004) and Salas et al. (2005) about the impact of the sampling time (and aggregation scale) on chaos detection and, more interestingly, it provides a possible explanation, in terms of information content, of the contrasting results available in the literature. Under the above mentioned conditions (sampling time and noise), samples drawn from chaotic and

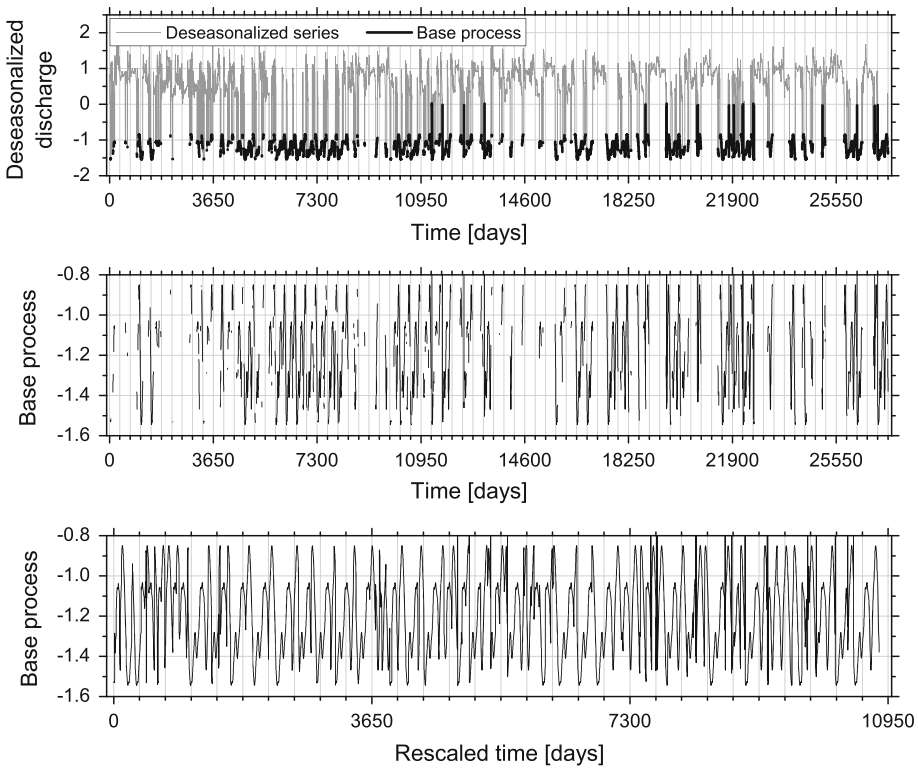

Fig. 13 Time series plots of the Nueces River deseasonalized data. The *top panel* shows the spurious base process (*black lines*) resulting by the deseasonalization of sequences of zero values (zero flow) and the superposed process resulting from the deseasonalization of positive stream flow values. The base process is magnified in the middle and *bottom panels*, by rescaling the y-axis and removing the time intervals corresponding to the superposed process

stochastic systems can contain similar amount of information (based on the quantifiers $\mathcal{H}_S$ and $\mathcal{C}_{JS}$). Thus, under the hypothesis that nonlinear low-dimensional dynamics are in act, it seems reasonable that their detection is possible only in some cases and is not so easy. In the CECP plane, a genuine and properly sampled chaotic signal would lie in the upper part of the admissible values, corresponding to the highest complexity values and mid-low entropy values. When the quantifiers fall far from this area, the study of the patterns of $\mathcal{H}_S$ and $\mathcal{C}_{JS}$ for different values of the delay time τ can help to shed light on possible oversampling (see De Micco et al. 2010; see Zunino et al. 2012b).

Our results contribute from an empirical point of view to the discussion about the observational equivalence between measure-theoretic deterministic systems and stochastic processes, in which observational equivalence means that deterministic description, when observed, and indeterministic descriptions gives the same predictions (Werndl 2009). Moreover, if the evidence equally supports a deterministic and a stochastic description, there is underdetermination concerning which description is preferable relative to evidence (Werndl 2012). Indeed, Werndl (2009) showed that every stochastic process is observationally equivalent to a deterministic system and many deterministic systems are observationally equivalent to stochastic processes. By studying a simple (non-linear deterministic) toy model describing a caricature hydrologic system, Koutsoyiannis (2010) showed that determinism and randomness coexist, and deciding which of them dominates depends on the horizon and scale of prediction. In this context, our results concerning the effect of the sampling frequency and the similarity of CECP patterns for deterministic and stochastic processes under particular conditions provide an empirical support to observational equivalence and its consequences.

The analysis of 80 long stream flow series from the continental US showed that the patterns described by $\mathcal{H}_S$ and $\mathcal{C}_{JS}$ for different values of the delay time τ allow highlighting important features of the signal information content such as its evolution with the sampling interval. Plotting the pairs $(\mathcal{H}_S, \mathcal{C}_{JS})$ for two characteristic delay times ($\tau = \{1,150\}$ days), the position of the complexity quantifiers for $\tau = 1$ pointed out that the stream flow series lie in an area that is shared by oversampled noisy chaotic/deterministic systems and seasonal stochastic signals. This ambiguity, which can be interpreted in terms of observational equivalence, can partially justify the mixed outcomes of the past attempts of distinguishing chaotic and stochastic behavior in real word geophysical signals from the point of view of the information contents. The comparison with the patterns described by surrogate series (drawn by different simulation techniques) showed that seasonal stochastic systems produce similar patterns, highlighting the discrepancies related to features such as zero flow sequences and intra-annual periodicities. This comparison also highlighted the effect of the

combination of the seasonality and long-term correlated residuals on the information content at different sampling time scales. Therefore, the CECP may also be used in perspective as a complementary tool to assess the adequacy of a simulation scheme in terms of information content at different time scales.

The comparison between the normalized entropy $\mathcal{H}_S$ and several stream flow attributes indicated that the degree of temporal organization (i.e., the temporal structure of dependence) of the signals is negatively correlated with the Hurst parameter H and the BFI values, whereas the opposite holds for $\mathcal{C}_{JS}$. Combining these results with those obtained by the pairwise analysis of other attributes, we can conclude that $\mathcal{C}_{JS}, H$ and BFI can be considered as proxies of the low-frequency groundwater processes that drive the long-term response of the watershed, whereas the drainage area is a proxy for high-frequency surface processes as is indicated by the strong relationships with the stream flow quantiles. It is worth noting that the relationship between the CECP quantifiers and the Hurst parameter is not surprising because analytical relationships do exist in the case of Hurst–Kolmogorov dynamics [fGn and the corresponding fBm; Bandt and Shiha (2007), Zunino et al. (2008)], and these processes can result from the extremization of entropy production of stationary stochastic representations of natural systems (Koutsoyiannis 2011b).

The comparison of some simple diagnostic plots based on linear and nonlinear indexes confirmed the importance of using "middle-ground" approaches (Sivakumar 2009) to obtain a more comprehensive description of the complexity that usually characterizes geophysical time series.

6 Conclusions

In this study, we have investigated the properties of a diagnostic plot called complexity–entropy causality plane based on the concepts of entropy and complexity, and the symbolization of the patterns retrieved by the phase space reconstruction. The method is nonparametric, meaning that the quantifiers that are used to draw the diagram depend on the ranks instead of the signed values of the analyzed time series. Under ideal conditions, linear or nonlinear, stochastic or chaotic deterministic dynamics can be discriminated in the CECP, even in presence of moderate observational noise, based on the position of the complexity quantifiers within the CECP area: chaotic deterministic systems tend to fall in the left region of the CECP and near the upper bound, while stochastic and highly noisy signals tend to occupy the right region close to the limit point (1, 0).

The CECP quantifiers depend on the delay time τ chosen for the phase space reconstruction. By changing the embedding delay we can change the sampling interval of the original system under analysis. Thus, the examination of the pattern of the CECP quantifiers for different τ values allows us to explore the temporal structure of the signal, highlighting periodic components and their information content. Moreover, the position of the complexity quantifiers in the CECP has a direct physical interpretation in terms of information content. Therefore, the CECP can be considered an example of "middle-ground" approach (Sivakumar 2009) that integrates different methods, deduced from stochastic and nonlinear dynamics analysis, in order to provide a balanced tool, which is less biased towards the two extreme, purely chaotic or stochastic points of view.

A few concluding remarks concern the implementation of the methodology and possible future lines of research. Bandt-Pompe symbolization method aims at alleviating the curse of dimensionality in estimating a continuous density function in the reconstructed state-space by replacing this density with the density of the ordinal patterns. Therefore, selecting $D = 6$ implies the use of 6! (720) rank based patterns instead of 10^6 data points required for continuous density estimation. With 27, 375 data points (the length of the analyzed stream flow series) the method approximately allocates 38 points per pattern. However, Figs. 6 and 7 show that the uncertainty of the estimates is negligible.

As mentioned in Sect. 2.2, the CECP quantifiers are devised for continuous variables, so that the ranks of the sampled observations are well defined. Therefore, the sequences of zeros of the intermittent rivers, as well as possible ties resulting from the instrument accuracy of some geophysical variables such as the rainfall, might affect the results. In this respect, Bian et al. (2012) proposed a modified permutation entropy estimator accounting for presence of ties. Even though preliminary analyses showed no substantial differences in the CECP results for our stream flow dataset, we are planning further analyses and a comparison of the estimators on zero-inflated rounded-off rainfall data.

Finally, the CECP quantifiers were computed on the complete time series to obtain an overall picture of the information content. However, since several rivers are characterized by regime changes related to human activities, a future direction of research can be the study of the evolution of CECP quantifiers on sliding windows (see e.g., Zunino et al. 2011) to quantify the change of information content due to regime shifts.

Acknowledgments Francesco Serinaldi acknowledges financial support from the Willis Research Network. Luciano Zunino and Osvaldo A. Rosso were supported by Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Argentina. Osvaldo A. Rosso gratefully acknowledges the support from FAPEAU fellowship, Brazil. The authors thank Dr. Bellie Sivakumar (University of New South Wales, Australia) for his useful remarks on an earlier

version of this paper, and two anonymous reviewers for their comments and suggestions. The analyses were performed in R (R Development Core Team 2009) with the help of the contributed packages `fractal` (Constantine and Percival 2007), `fArma` (Wuertz et al. 2008), `tseriesChaos` (Di Narzo 2007), `tsDyn` (Di Narzo and Aznarte 2007) and `msProcess` (Gong et al. 2009). The authors and maintainers of this software are gratefully acknowledged.

Appendix

The equations of the chaotic systems used in this study are listed as follows. The sets of parameters were chosen so that the systems describe chaotic attractors.

Duffing system:

$$\begin{cases} \dot{x} = y \\ \dot{y} = -x^3 - cy + F\cos(z), \\ \dot{z} = \Omega \end{cases} \tag{6}$$

where $c = 0.05$, $F = 7.5$ and $\Omega = 1$.

Lorenz system:

$$\begin{cases} \dot{x} = a(y - x) \\ \dot{y} = x(b - z) - y, \\ \dot{z} = xy + cz \end{cases} \tag{7}$$

where $a = 10$, $b = 28$ and $c = -8/3$.

Rossler system:

$$\begin{cases} \dot{x} = -(y + z) \\ \dot{y} = x + ay \\ \dot{z} = b + z(x - c) \end{cases}, \tag{8}$$

where $a = 0.2$, $b = 0.2$ and $c = 5.7$.

Henon map:

$$\begin{cases} x_t = a - x_{t-1}^2 + by_{t-1} \\ y_t = x_{t-1} \end{cases}, \tag{9}$$

where $a = 1.4$ and $b = 0.3$.

References

Bandt C, Pompe B (2002) Permutation entropy: a natural complexity measure for time series. Phys Rev Lett 88(7):174102

Bandt C, Shiha F (2007) Order patterns in time series. J Time Ser Anal 28(5):646–665

Beran J (1989) A test of location for data with slowly decaying serial correlations. Biometrika 76(2):261–269

Beran J (1994) Statistics for long-memory processes. Chapman & Hall, London

Bian C, Qin C, Ma QDY, Shen Q (2012) Modified permutation–entropy analysis of heartbeat dynamics. Phys Rev E 85:021906

Bonneville Power Administration, US Bureau of Reclamation, US Army Corps of Engineers (2001) The Columbia River System: inside story, 2nd edn. Bonneville Power Administration, Portland, Oregon. http://www.bpa.gov/corporate/Power_of_Learning/docs/columbia_river_inside_story.pdf

Cánovas JS, Guillamón A, delCarmen Ruíz M (2011) Using permutations to detect dependence between time series. Phys D 240(14–15):1199–1204

Carpi LC, Saco PM, Rosso OA (2010) Missing ordinal patterns in correlated noises. Phys A 389(10):2020–2029

Castellarin A, Burn DH, Brath A (2001) Assessing the effectiveness of hydrological similarity measures for flood frequency analysis. J Hydrol 241(3–4):270–285

Chaitin GJ (1966) On the length of programs for computing finite binary sequences. J Assoc Comput Mach 13(4):547–569

Cleveland WS, Devlin SJ (1988) Locally-weighted regression: an approach to regression analysis by local fitting. J Am Stat Assoc 83(403):596–610

Constantine W, Percival D (2007) Fractal: insightful fractal time series modeling and analysis. R package version 1.0-2

Crutchfield JP, Young K (1989) Inferring statistical complexity. Phys Rev Lett 63(2):105–108

De Micco L, Larrondo HA, Plastino A, Rosso OA (2009) Quantifiers for randomness of chaotic pseudo-random number generators. Philos Trans R Soc A 367(1901):3281–3296

De Micco L, Fernández JG, Larrondo HA, Plastino A, Rosso OA (2010) Sampling period, statistical complexity, and chaotic attractors. Phys A 391(8):2564–2575

Di Narzo AF (2007) TseriesChaos: analysis of nonlinear time series. R package version 0.1-8

Di Narzo AF, Aznarte JL (2007) tsDyn: time series analysis based on dynamical systems theory. R package version 0.6

Dooge JCI (1968) The hydrologic system as a closed system. Bull Int Assoc Sci Hydrol 13(1):58–68

Escalona-Morán M, Cosenza MG, López-Ruiz R, García P (2010) Statistical complexity and nontrivial collective behavior in electroencephalografic signals. Int J Bifurcat Chaos 20(6):1723–1729

Gong L, Constantine W, Chen YA (2009) msProcess: protein mass spectra processing. http://www.insightful.com/services/research/proteome/default.asp . R package version 1.0.5

Grassberger P (1986) Toward a quantitative theory of self-generated complexity. Int J Theor Phys 25(9):907–938

Grassberger P, Procaccia I (1983) Characterization of strange attractors. Phys Rev Lett 50(5):346–349

Grayson RB, Blöschl G (2000) Spatial patterns in catchment hydrology: observations and modeling. Cambridge University Press, Cambridge

Grosse I, Bernaola-Galván P, Carpena P, Román-Roldán R, Oliver J, Stanley HE (2002) Analysis of symbolic sequences using the Jensen–Shannon divergence. Phys Rev E 65(4):041905

Hauhs M, Lange H (2008) Classification of runoff in headwater catchments: a physical problem. Geogr Compass 2(1):235–254

Higuchi T (1988) Approach to an irregular time series on the basis of the fractal theory. Phys D 31(2):277–283

Hirpa FA, Gebremichael M, Over TM (2010) River flow fluctuation analysis: effect of watershed area. Water Resour Res 46(12):W12529

Hu K, Ivanov PC, Chen Z, Carpena P, Stanley HE (2001) Effect of trends on detrended fluctuation analysis. Phys Rev E 64(1):011114

Kantelhardt JW, Zschiegner SA, Koscielny-Bunde E, Havlin S, Bunde A, Stanley HE (2002) Multifractal detrended fluctuation analysis of nonstationary time series. Phys A 316(1–4):87–114

Kantelhardt JW, Rybski D, Zschiegner SA, Braun P, Koscielny-Bunde E, Livina V, Havlin S, Bunde A (2003) Multifractality of river runoff and precipitation: comparison of fluctuation analysis and wavelet methods. Phys A 330(1–2):240–245

Kantelhardt JW, Koscielny-Bunde E, Rybski D, Braun P, Bunde A, Havlin S (2006) Long-term persistence and multifractality of

precipitation and river runoff records. J Geophys Res 111(D1):D01106
Király A, Jánosi IM (1998) Stochastic modeling of daily temperature fluctuations. Phys Rev E 65(5):051102
Kolmogorov AN (1965) Three approaches to the quantitative definition of information. Probl Inf Transm 1:1–7
Koscielny-Bunde E, Kantelhardt JW, Braun P, Bunde A, Havlin S (2006) Long-term persistence and multifractality of river runoff records: detrended fluctuation studies. J Hydrol 322(1–4):120–137
Koutsoyiannis D (2010) HESS Opinions "A random walk on water". Hydrol Earth Syst Sci 14(3):585–601
Koutsoyiannis D (2011a) Hurst–Kolmogorov dynamics and uncertainty. J Am Water Resour Assoc 47(3):481–495
Koutsoyiannis D (2011b) Hurst–Kolmogorov dynamics as a result of extremal entropy production. Phys A 390(8):1424–1432
Kowalski A, Martín MT, Plastino A, Rosso OA (2007) Bandt–Pompe approach to the classical-quantum transition. Phys D 233(1):21–31
Krasovskaia I (1995) Quantification of the stability of river flow regimes. Hydrol Sci J 40(5):587–598
Krasovskaia I (1997) Entropy-based grouping of river flow regimes. J Hydrol 202(1–4):173–191
Lamberti PW, Martín MT, Plastino A, Rosso OA (2004) Intensive entropic non-triviality measure. Phys A 334(1–2):119–131
Lange H, Rosso OA, Hauhs M (2013) Ordinal pattern and statistical complexity analysis of daily stream flow time series. Eur Phys J Spec Top 222(2):535–552
Lempel A, Ziv J (1976) On the complexity of finite sequences. IEEE Trans Inf Theory 22(1):75–81
Livina V, Ashkenazy Y, Kizner Z, Strygin V, Bunde A, Havlin S (2003) A stochastic model of river discharge fluctuations. Phys A 330(1–2):283–290
López-Ruiz R, Mancini HL, Calbet X (1995) A statistical measure of complexity. Phys Lett A 209(5–6):321–326
López-Ruiz R, Sañudo J, Romera E, Calbet X (2011) Statistical complexity and Fisher–Shannon information: applications. In: Sen KD (ed) Statistical complexity. Springer, Netherlands, pp 65–127
Ludescher J, Bogachev MI, Kantelhardt JW, Schumann AY, Bunde A (2011) On spurious and corrupted multifractality: the effects of additive noise, short-term memory and periodic trends. Phys A 390(13):2480–2490
Marković D, Koch M (2005) Sensitivity of Hurst parameter estimation to periodic signals in time series and filtering approaches. Geophys Res Lett 32(17):L17401
Martín MT, Plastino A, Rosso OA (2003) Statistical complexity and disequilibrium. Phys Lett A 311(2–3):126–132
Martín MT, Plastino A, Rosso OA (2006) Generalized statistical complexity measures: geometrical and analytical properties. Phys A 369(2):439–462
McDonnell JJ, Woods RA (2004) On the need for catchment classification. J Hydrol 299(1–2):2–3
Montanari A (2005) Deseasonalisation of hydrological time series through the normal quantile transform. J Hydrol 313(3–4): 274–282
Montanari A, Rosso R, Taqqu MS (1997) Fractionally differenced ARIMA models applied to hydrologic time series: identification, estimation, and simulation. Water Resour Res 33(5):1035–1044
Montanari A, Rosso R, Taqqu MS (2000) A seasonal fractional ARIMA model applied to the Nile River monthly flows at Aswan. Water Resour Res 36(5):1249–1259
Movahed MS, Hermanis E (2008) Fractal analysis of river flow fluctuations. Phys A 387(4):915–932
Mudelsee M (2007) Long memory of rivers from spatial aggregation. Water Resour Res 43(1):W01202
Pachepsky Y, Guber A, Jacques D, Simunek J, Genuchten MTV, Nicholson T, Cady R (2006) Information content and complexity of simulated soil water fluxes. Geoderma 134(3–4):253–266
Pan F, Pachepsky YA, Guber AK, Hill RL (2011) Information and complexity measures applied to observed and simulated soil moisture time series. Hydrol Sci J 56(6):1027–1039
Pan F, Pachepsky YA, Guber AK, McPherson BJ, Hill RL (2012) Scale effects on information theory-based measures applied to streamflow patterns in two rural watersheds. J Hydrol 414–415:99–107
Peng CK, Buldyrev SV, Havlin S, Simons M, Stanley HE, Goldberger AL (1994) Mosaic organization of DNA nucleotides. Phys Rev E 49(2):1685–1689
R Development Core Team (2009) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org, ISBN 3-900051-07-0
Rego CRC, Frota HO, Gusmão MS (2013) Multifractality of Brazilian rivers. J Hydrol 495:208–215
Regonda SK, Sivakumar B, Jain A (2004) Temporal scaling in river flow: can it be chaotic. Hydrol Sci J 49(3):373–385
Rosso OA, Masoller C (2009a) Detecting and quantifying stochastic and coherence resonances via information-theory complexity measurements. Phys Rev E 79(4):040106(R)
Rosso OA, Masoller C (2009b) Detecting and quantifying temporal correlations in stochastic resonance via information theory measures. Eur Phys J B 69(1):37–43
Rosso OA, Larrondo HA, Martín MT, Plastino A, Fuentes MA (2007a) Distinguishing noise from chaos. Phys Rev Lett 99(15): 154102
Rosso OA, Zunino L, Pérez DG, Figliola A, Larrondo HA, Garavaglia M, Martín MT, Plastino A (2007b) Extracting features of Gaussian self-similar stochastic processes via the Bandt–Pompe approach. Phys Rev E 76(6):061114
Rosso OA, Carpi LC, Saco PM, Ravetti MG, Larrondo HA, Plastino A (2012a) The Amigó paradigm of forbidden/missing patterns: a detailed analysis. Eur Phys J B 85(12):419
Rosso OA, Carpi LC, Saco PM, Ravetti MG, Plastino A, Larrondo HA (2012b) Causality and the entropy–complexity plane: robustness and missing ordinal patterns. Phys A 391(1–2):42–55
Rosso OA, Olivares F, Zunino L, Micco L, Aquino ALL, Plastino A, Larrondo HA (2013) Characterization of chaotic maps using the permutation Bandt–Pompe probability distribution. Eur Phys J B 86(4):116
Salas JD, Kim HS, Eykholt R, Burlando P, Green TR (2005) Aggregation and sampling in deterministic chaos: implications for chaos identification in hydrological processes. Nonlinear Process Geophys 12(4):557–567
Sánchez JR, López-Ruiz R (2005) A method to discern complexity in two-dimensional patterns generated by coupled map lattices. Phys A 355(2–4):633–640
Serinaldi F (2010) Use and misuse of some Hurst parameter estimators applied to stationary and non-stationary financial time series. Phys A 389(14):2770–2781
Sivakumar B (2004) Dominant processes concept in hydrology: moving forward. Hydrol Process 18(12):2349–2353
Sivakumar B (2008) Dominant processes concept, model simplification and classification framework in catchment hydrology. Stoch Environ Res Risk Assess 22(6):737–748
Sivakumar B (2009) Nonlinear dynamics and chaos in hydrologic systems: latest developments and a look forward. Stoch Environ Res Risk Assess 23(7):1027–1036
Sivakumar B, Singh VP (2012) Hydrologic system complexity and nonlinear dynamic concepts for a catchment classification framework. Hydrol Earth Syst Sci 16(11):4119–4131
Sivakumar B, Jayawardena AW, Li WK (2007) Hydrologic complexity and classification: a simple data reconstruction approach. Hydrol Process 21(20):2713–2728
Smakhtin VY (2001) Low flow hydrology: a review. J Hydrol 240(3–4):147–186

Soriano MC, Zunino L, Larger L, Fischer I, Mirasso CR (2011a) Distinguishing fingerprints of hyperchaotic and stochastic dynamics in optical chaos from a delayed opto-electronic oscillator. Opt Lett 36(12):2212–2214

Soriano MC, Zunino L, Rosso OA, Fischer I, Mirasso CR (2011b) Time scales of a chaotic semiconductor laser with optical feedback under the lens of a permutation information analysis. IEEE J Quantum Electron 47(2):252–261

Staniek M, Lehnertz K (2007) Parameter selection for permutation entropy measurements. Int J Bifurcat Chaos 17(10):3729–3733

Theiler J, Eubank S, Longtin A, Galdrikian B, Farmer JD (1992) Testing for nonlinearity in time series: the method of surrogate data. Phys D 58(1–4):77–94

Tiana-Alsina J, Torrent MC, Rosso OA, Masoller C, García-Ojalvo J (2010) Quantifying the statistical complexity of low-frequency fluctuations in semiconductor lasers with optical feedback. Phys Rev A 82(1):013,819

Tyralis H, Koutsoyiannis D (2011) Simultaneous estimation of the parameters of the Hurst–Kolmogorov stochastic process. Stoch Environ Res Risk Assess 25(1):21–33

Wackerbauer R, Witt A, Atmanspacher H, Kurths J, Scheingraber H (1994) A comparative classification of complexity measures. Chaos Solitons Fractals 4(1):133–173

Wang W, Van Gelder PHAJM, Vrijling JK, Chen X (2007) Detecting long-memory: Monte carlo simulations and application to daily streamflow processes. Hydrol Earth Syst Sci 11(2):851–862

Werndl C (2009) Are deterministic descriptions and indeterministic descriptions observationally equivalent. Stud Hist Philos Sci B 40(3):232–242

Werndl C (2012) Evidence for the deterministic or the indeterministic description? A critique of the literature about classical dynamical systems. J Gen Philos Sci 43(2):295–312

Wuertz D et al (2008) fArma: ARMA time series modelling. http://www.rmetrics.org, R package version 270.74

Zanin M, Zunino L, Rosso OA, Papo D (2012) Permutation entropy and its main biomedical and econophysics applications: a review. Entropy 14(8):1553–1577

Zhang Q, Zhou Y, Singh VP, Chen YD (2011) Comparison of detrending methods for fluctuation analysis in hydrology. J Hydrol 400(1–2):121–132

Zunino L, Pérez DG, Martín MT, Garavaglia M, Plastino A, Rosso OA (2008) Permutation entropy of fractional Brownian motion and fractional Gaussian noise. Phys Lett A 372(27–28):4768–4774

Zunino L, Zanin M, Tabak BM, Pérez DG, Rosso OA (2009) Forbidden patterns, permutation entropy and stock market inefficiency. Phys A 388(14):2854–2864

Zunino L, Soriano MC, Fischer I, Rosso OA, Mirasso CR (2010a) Permutation information theory approach to unveil delay dynamics from time series analysis. Phys Rev E 82(4):046,212

Zunino L, Zanin M, Tabak BM, Pérez DG, Rosso OA (2010b) Complexity–entropy causality plane: a useful approach to quantify the stock market inefficiency. Phys A 389(9):1891–1901

Zunino L, Tabak BM, Serinaldi F, Zanin M, Pérez DG, Rosso OA (2011) Commodity predictability analysis with a permutation information theory approach. Phys A 390(5):876–890

Zunino L, Fernández Bariviera A, Guercio MB, Martinez LB, Rosso OA (2012a) On the efficiency of sovereign bond markets. Phys A 391(18):4342–4349

Zunino L, Soriano M, Rosso O (2012b) Distinguishing chaotic and stochastic dynamics from time series by using a multiscale symbolic approach. Phys Rev E Stat Nonlin Soft Matter Phys 86(4):046210