Eur. Phys. J. Special Topics **222**, 535–552 (2013)

DOI: 10.1140/epjst/e2013-01858-3

THE EUROPEAN PHYSICAL JOURNAL SPECIAL TOPICS

Regular Article

# Ordinal pattern and statistical complexity analysis of daily stream flow time series

H. Lange[1,a], O.A. Rosso[2,3], and M. Hauhs[4]

[1] Norwegian Forest and Landscape Institute, Postsboks 115, N-1430 Ås, Norway
[2] LaCCAN/CPMAT, Instituto de Computação, Universidade Federal de Alagoas BR 104 Norte km 97, 57072-970 Maceió, Alagoas, Brazil
[3] Laboratorio de Sistemas Complejos, Facultad de Ingeniería, Universidad de Buenos Aires, 1063 Av. Paseo Colón 840, Ciudad Autónoma de Buenos Aires, Argentina
[4] Ecological Modelling, University of Bayreuth, 95440 Bayreuth, Germany

Received 27 March 2013 / Received in final form 25 April 2013
Published online 25 June 2013

**Abstract.** When calculating the Bandt and Pompe ordinal pattern distribution from given time series at depth $D$, some of the $D!$ patterns might not appear. This could be a pure finite size effect (missing patterns) or due to dynamical properties of the observed system (forbidden patterns). For pure noise, no forbidden patterns occur, contrary to deterministic chaotic maps. We investigate long time series of river runoff for missing patterns and calculate two global properties of their pattern distributions: the Permutation Entropy and the Permutation Statistical Complexity. This is compared to purely stochastic but long-range correlated processes, the $k$-noise (noise with power spectrum $f^{-k}$), where $k$ is a parameter determining the strength of the correlations. Although these processes closely resemble runoff series in their correlation behavior, the ordinal pattern statistics reveals qualitative differences, which can be phrased in terms of missing patterns behavior or the temporal asymmetry of the observed series. For the latter, an index is developed in the paper, which may be used to quantify the asymmetry of natural processes as opposed to artificially generated data.

## 1 Introduction

Hydrological catchments, the basins of attraction of water running through given points in a landscape, are the basic spatial units considered in hydrology. They are conceptualized as complex dynamical systems where deterministic and stochastic processes occur simultaneously. It is increasingly recognized that there is a need for the classification of catchments and hydrological phenomena [1–4]. This is in part driven by modeling requirements, where the identification of the appropriate model type and complexity and data requirements is key, and in part by attempts to

[a] e-mail: holger.lange@skogoglandskap.no

transfer information gained from fully equipped catchments (gauged catchments) to ones where runoff measurements are lacking (ungauged catchments). The classification is a crucial step towards a unified theory of catchment dynamics; however, the task is far from being straightforward. Categories which might be used in the unifying approach refer to the set of dominant processes, soil types, geology, climate, or vegetation cover, focussing on the flow paths or the *structure* of the catchment. Alternatively, tracer studies might be used to get insight into travel time distributions, focussing on the temporal dynamics or the *function* of the catchment in light of its history. There are strong hints that biological activity is shaping the evolution of river networks and thus land surfaces [5].

Our approach for classification is strictly data-driven. We do not make any explicit assumptions about relevant hydrological processes. Due to the relevance of water resources for human civilizations, a large number of runoff records from hydrological catchments exists as a result of environmental monitoring. These long-term time series, typically at daily resolution and extending over several decades, form the raw material of the analysis presented here. Since we perform time series analysis, the focus is necessarily on the temporal (functional) aspects of the system. However, a next step towards a new classification will be relating these to system attributes referring to climate, soils, geology, vegetation cover, natural history and other aspects of the respective catchments. Thus, we are seeking to combine the two classification approaches just mentioned.

We consider a set of almost 500 individual streamflow time series where the catchment size is varying over several orders of magnitude. As a consequence, the magnitude of the signal is also strongly varying – given comparable precipitation amounts, the long-term water balance requires that cumulative runoff is linear proportional to the catchment size. It is thus convenient to express runoff in mm instead ($1\,\mathrm{mm} = 1\,\mathrm{million\ litres/km^2}$). This requires, however, that the catchment size is known, a potential problem for ungauged systems. The scale of the measurements is irrelevant for our purposes, and we are seeking a robust, universally applicable approach.

The analysis method selected depends on several properties of the data sets, in particular their length, the presence of gaps, outliers, trends and other instationarities. For sufficiently long time series, using order statistics [6,7] is among the obvious and promising choices. Thus, we convert the time series to order patterns at a fixed depth $D$. Taking into account the typical lengths of our time series (several thousands of observations), a recommended value for the depth is $D = 6$. The distribution of order patterns will be investigated in detail, with a special focus on missing patterns and temporal asymmetry. Since we are interested in a separation of deterministic and stochastic parts of these time series, we compare the results to a reference stochastic process, the $k$ noise, and to a set of artificial deterministic-chaotic systems (iterated maps) free of noise. From the order pattern distributions, we calculate two indices, their permutation entropy and the permutation statistical complexity [7–9]. The former quantifies the information content of the distribution, whereas the latter describes its complexity. They can be combined in a two-dimensional diagram known as the Complexity-Entropy Causality Plane, or CECP [9,10]. Each time series is represented as a point in this diagram, and a visual comparison of runoff, $k$ noise and deterministic chaos is particularly easy and illustrative.

The paper is organized as follows. Section 2 contains a description of the data used and the artificial reference processes. The next Sect. 3 gives the details of our methodological approach. Section 4 presents selected results for individual runoff series, and the CECP plane. We put these results into the context of hydrological modeling and provide an outlook in the final Sect. 5.

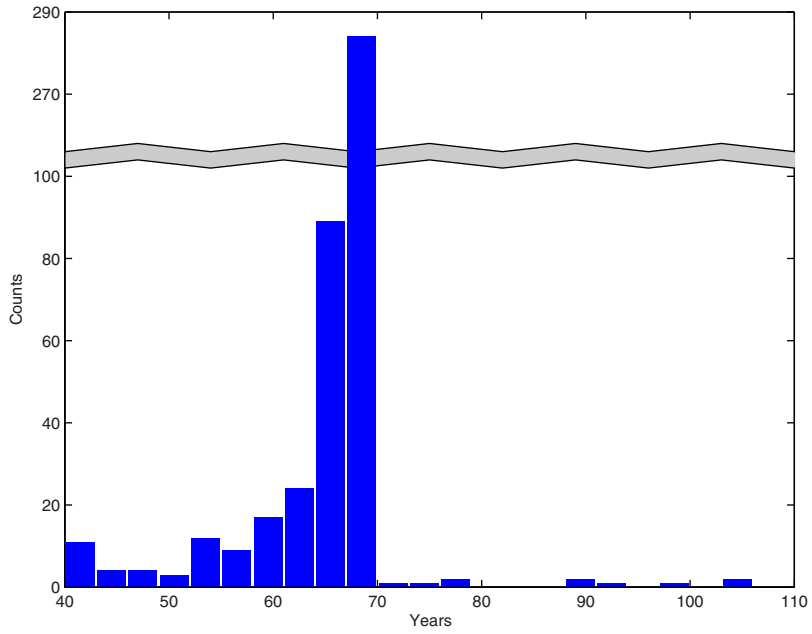


**Fig. 1.** The histogram of the length of all runoff time series in years. Note that the y axis is interrupted; the actual value of the bin counts close to 70 years is 284.

## 2 Data sets and artificial processes

Runoff time series at daily resolution are available from a number of archives. We would like to mention the Global Runoff Data Centre (GRDC[1]) in Koblenz, Germany, and the United States Geological Survey (USGS) water information system[2]. From the latter, we used 438 time series covering the continental United States. Due to personal contacts (cf. Acknowledgments), we received 14 series from South America (mainly Argentina) from the Subsecretaría de Recursos Hídricos[3], and 48 series from New Zealand from the National Institute of Water and Atmospheric Research[4].

A notorious problem of long-term hydrological records (and environmental observations in general) is the presence of missing values or gaps. It is not our intention in the current investigation to deal with the artefacts induced by gap-filling algorithms. Thus, we chose strictly gap-free series.

The lengths of the time series vary between 40 and 105 years (more than 38000 values) with a mean of 66 years. In Fig. 1, a histogram of the lengths is provided. The majority of time series is from a compilation of US-american rivers extending from 1938 to 2008, which explains the peak at 70 years. Globally, only very few uninterrupted daily records with a length $> 100$ years are available. These values have an implication for the choice of the embedding dimension (cf. Methods section). The selection covers a wide range of catchment sizes, from less than $10\,\mathrm{km}^2$ to more than $10^6\,\mathrm{km}^2$, average annual rainfall, annual temperature, geological conditions and so on. Thus, if these factors are important for the dynamics of the runoff as reflected

[1] http://www.bafg.de/GRDC/
[2] http://waterdata.usgs.gov/nwis
[3] http://www.hidricosargentina.gov.ar/
[4] http://www.niwa.co.nz/

in the order statistics of the time series, we can expect contrasts in the indicators calculated here.

We expect the time series to be complicated mixtures of deterministic and stochastic parts. They are strongly autocorrelated – the lag-1 autocorrelation coefficient usually exceeds 0.9 for daily data. Thus, uncorrelated white noise cannot be expected to be a meaningful model process for the stochastic part. Instead, we are using long-range correlated noise with the power spectra decaying only as a power law of the frequency, the $k$-noise, as a reference process. The short-term decay of the autocorrelation function for the individual runoff time series can then be mimicked by tuning the value of $k$. However, since we are by necessity investigating short-time dynamics only, where the relevant time scale is given by the embedding dimension, agreement in the dynamics of runoff and $k$-noise in that temporal regime does not imply that the long-term structure of the former is reflected in the latter. We also expect a set of long-term periodic components in the natural record driven by atmospheric forcings; these are of course absent from the reference noise process.

## 3 Methods

The starting point of our study are observed time series from natural systems, i.e. given sets $\mathcal{X} \equiv \{x_t, t = 1, \cdots, N\}$, with $N$ being the number of observations. These time series often appear irregular, highly fluctuating and difficult to predict. Part of that behaviour is due to noise in the system or the measurement process. However, the concept of low dimensional deterministic chaos has shown that nonlinearities, albeit strictly deterministic, may lead to behaviour which also seems difficult to predict, indistinguishable from noisy series at first sight.

Signals emerging from chaotic systems occupy a place intermediate between predictable regular or quasi-periodic signals and totally irregular stochastic signals (noise) which are completely unpredictable. Chaotic time series are irregular in time, barely predictable, and exhibit interesting structures in the phase space. They are representatives of a set of signals exhibiting complex non-periodic behaviour with continuous, broad band Fourier spectra. Chaotic systems display "sensitivity to initial conditions" which are the cause of instability everywhere in phase-space. These instabilities uncover information about the phase-space "population", not available otherwise [11]. In turn this leads us to think of chaos as an *information source,* whose associated rate of generated information is formulated in precise fashion via the Kolmogorov-Sinai entropy [12,13].

These considerations motivate our present interest in the computation of quantifiers based on Information Theory, like "entropy", "statistical complexity", "entropy-complexity plane", etc. These quantifiers can be used to detect determinism in time series [9]. Indeed, different Information Theory based measures (normalized Shannon entropy and statistical complexity) allow for a better distinction between deterministic chaotic and stochastic dynamics whenever "causal" information is incorporated via the Bandt and Pompe (BP) methodology [6]. For a review of BP's methodology and its applications to physics, biomedical and econophysic signals, see [14]. In the following, we provide a technical description of the quantifiers we are using for the hydrological time series investigated.

### 3.1 Information theory quantifiers

As is customary since Shannon's seminal work [15], we discuss the information content of data in terms of information-theoretic entropy. As a first step, a discretization of

the phase space is required to obtain a *finite* set of possibles *states* of the system. The information content of the system then is typically evaluated from its probability distribution function (PDF). Consider a given time series $\mathcal{X} = \{x_t, t = 1, \cdots, N\}$ of length $N$ and the corresponding discrete probability distribution set $P = \{p_i;\ i = 1, \cdots, M\}$ describing the empirical distribution of $\mathcal{X}$, in which $M$ is the number of possible states. Shannon's logarithmic information measure reads [15]

$$S[P] \ = \ -\sum_{i=1}^{M} \ p_i \ \ln(p_i) \ . \tag{1}$$

The Shannon entropy $S$ is regarded as a measure of the uncertainty associated to the physical processes described by the probability distribution $P$. From now on we assume that the only restriction on the PDF representing the state of our system is $\sum_{j=1}^{M} p_j = 1$ (micro-canonical representation in statistical mechanics terms). If $S[P] = 0$, there is only one non-vanishing $p_i$, and we know the state of the system with certainty. Our knowledge of the underlying process described by the probability distribution is in this instance maximal. On the other hand, our ignorance is maximal for a uniform distribution, $P_e = \{1/M, \cdots, 1/M\}$, and in this case $S[P_e] = S_{max} = \ln(M)$. Since predictability of the system behaviour is minimal in this case, apparent randomness is at a maximum. The Shannon entropy is a quantifier for randomness.

There is no universally accepted definition of *complexity*. Intuitively, complexity should be related to the amount of structure or the number of patterns present in a system. One would like to have some functional $\mathcal{C}[P]$ adequately capturing the "structuredness" in the same way as Shannon's entropy [15] captures randomness.

It is widely known that an entropic measure does not quantify the degree of structure or patterns present in a process [16]. Moreover, it was recently shown that measures of statistical or structural complexity are necessary for a better understanding of chaotic time series because they are able to capture their organizational properties [17]. This specific kind of information is not revealed by randomness measures but needs an appropriate complexity quantifier.

Rosso and coworkers introduced an effective statistical complexity measure (SCM) that is able to *a)* detect essential details of the dynamics and *b)* differentiate between chaos and (different degrees of) periodicity [18]. This specific SCM provides important additional information regarding the peculiarities of the underlying PDF, not already detected by the entropy.

The thermodynamically intensive SCM is defined, following the intuitive notion advanced by López-Ruiz et al. [19], via the product

$$\mathcal{C}[P] = \mathcal{Q}_J[P, P_e] \cdot \mathcal{H}[P] \tag{2}$$

of *i)* the *normalized* Shannon entropy

$$\mathcal{H}[P] = \mathrm{S}[P]/\mathrm{S}_{max} \ , \tag{3}$$

with $\mathrm{S}_{max} = \mathrm{S}[P_e] = \ln M$, $(0 \leq \mathcal{H}_S \leq 1)$ and $P_e = \{1/M, \cdots, 1/M\}$ (the uniform distribution) and *ii)* the so-called disequilibrium $\mathcal{Q}_J$. This quantifier is defined in terms of the extensive (in the thermodynamical sense) Jensen-Shannon divergence $\mathcal{J}[P, P_e]$ that links two PDFs. We have

$$\mathcal{Q}_J[P, P_e] = Q_0 \cdot \mathcal{J}[P, P_e] \ , \tag{4}$$

with

$$\mathcal{J}[P, P_e] = \mathrm{S}\left[(P + P_e)/2\right] - \mathrm{S}[P]/2 - \mathrm{S}[P_e]/2 \ . \tag{5}$$

$Q_0$ is a normalization constant ($0 \leq \mathcal{Q}_J \leq 1$), equal to the inverse of the maximum possible value of $\mathcal{J}[P, P_e]$. This value is obtained when one of the values of $P$, say $p_m$, is equal to one and the remaining $p_i$ values are equal to zero. Note that in this case, the entropy is zero, and thus the complexity (see Eq. (2)). Explicit expressions for the $p_i$ values which maximize the complexity $\mathcal{C}$ for prescribed values of $\mathcal{H}$ are provided in [8].

The Jensen-Shannon divergence that quantifies the difference between two (or more) probability distributions is especially useful to compare the symbol-composition of sequences [20]. The SCM constructed in this way has the intensive property found in many thermodynamic quantities [18]. We stress the fact that the statistical complexity defined above is the product of two normalized entropies (the Shannon entropy and Jensen-Shannon divergence), but it is a nontrivial function of the entropy because it depends on two different probability distributions, i.e., the one corresponding to the state of the system, $P$, and the uniform distribution, $P_e$, taken as reference distribution. A class of SCMs might be constructed by considering other reference distributions, e.g. from another time series to determine the Jensen-Shannon divergence between different observables from the same system, or the same observable obtained from different systems.

The temporal evolution of the intensive SCM can be analyzed using a diagram of $\mathcal{C}$ versus time $t$. The second law of thermodynamics states that for isolated systems entropy grows monotonically with time ($d\mathcal{H}/dt \geq 0$) [21]. This implies that $\mathcal{H}$ can be regarded as an arrow of time, so that an equivalent way to study the temporal evolution of the intensive SCM is through the analysis of $\mathcal{C}$ versus $\mathcal{H}$. In this way, the normalized entropy-axis substitutes for the time-axis. Furthermore, it has been shown that for a given value of $\mathcal{H}$, the range of possible statistical complexity values varies between a minimum $\mathcal{C}_{min}$ and a maximum $\mathcal{C}_{max}$ [8], restricting the possible values of the intensive SCM in this plane.

Therefore, calculating complexity provides additional insights into the details of the system's probability distribution, which is not discriminated by randomness measures like the entropy [9,17]. Complexity can also help to uncover information related to the correlational structures related to the components of the physical process under study [22,23]. The entropy-complexity diagram (or plane), $\mathcal{H} \times \mathcal{C}$, has been used to study changes in the dynamics of a system originating in modifications of some characteristic parameters (see, for instance, Ref. [14] and references therein).

### 3.2 Bandt-Pompe probability distribution

An important point for the evaluation of the previous information measurements is the proper determination of the underlying probability distribution function (PDF) $P$, associated to a given dynamical system or time series. Using the value distribution directly is not an option for time series since the dynamical aspects (e.g. autocorrelations) are completely ignored. Very different dynamical systems which happen to possess the same value PDF would be indistinguishable.

Bandt and Pompe (BP) introduced a by now very successful methodology in 2002 for the evaluation of the PDF associated to scalar time-series data using a symbolization technique [6]. The time series is converted to a symbolic series by first choosing an *embedding dimension* $D$ and a time lag $\tau$. The series is decimated by taking only one value per $\tau$ (for $\tau = 1$, the whole series is kept). The remaining numerical values in each window of length $D$ are then ranked in ascending order. These ranks, i.e. just the natural numbers from 1 to $D$, are the ingredients of the resulting *order* time series. This is tantamount to a phase space reconstruction with embedding dimension (pattern length) $D$ and delay $\tau$. For a precise definition and methodological details,

see [6,14]. As a result, the time series is represented by a sequence of permutations of $\{1,\ldots,D\}$, or *patterns*. Counting the number of occurrences of each of the $D!$ patterns gives the Bandt-Pompe PDF. For this choice of PDF, the $M$ appearing in Eq. (1) is simply $M = D!$.

Note that the symbol sequence arises naturally from the time series and no model-based assumptions are needed. This technique, as opposed to most of those currently used, takes into account the temporal structure (time causality) of the time series generated by the physical process under study. This feature allows us to uncover important details concerning the ordinal structure of the time series [7,9,24], and can also yield information about temporal correlation [22,23].

Ordinal time-series analysis entails losing some details of the original time-series amplitude-information. Nevertheless, the symbolic representation of time-series allows for an accurate empirical reconstruction of the underlying phase space, even in the presence of weak (observational and dynamical) noise [6]. Nonlinear drifts or scalings artificially introduced by a measurement device will not modify the quantifiers estimation (see, e.g., [25]).

Additional advantages of the method reside in *i)* its simplicity (we need few parameters: the pattern length/ embedding dimension $D$ and the embedding delay $\tau$) and *ii)* the extremely fast nature of the pertinent calculation-process [26,27]. The BP methodology can be applied not only to time series representative of low dimensional dynamical systems but also to any type of time series (regular, chaotic, noisy, experimentally obtained or artificially generated). In fact, the existence of an attractor in the $D$-dimensional phase space is not assumed. The only condition for the applicability of the BP methodology is a very weak stationary assumption: that is, for $k \leq D$, the probability for $x_t < x_{t+k}$ should not depend on $t$ [6].

The BP-generated probability distribution $P$ is obtained once we choose the embedding dimension $D$ and the embedding delay $\tau$. It has been established that the length $N$ of the time series must satisfy the condition $N \gg D!$ in order to achieve a reliable statistics and proper distinction between stochastic and deterministic dynamics [9]. With respect to the selection of the parameters, Bandt and Pompe suggest in their cornerstone paper [6] to work with $3 \leq D \leq 7$ with a time lag $\tau = 1$, a recommendation which we follow also in this work. At $D = 6$, we have $N > 10D!$ for all series which leads to reliable estimates of the asymptotic Bandt-Pompe PDF from our finite data. We also stick to $\tau = 1$. Nevertheless, other values of $\tau$ might provide additional information. Soriano et al. [28,29] and Zunino et al. [30,31], recently, showed that this parameter is strongly related, when it is relevant, to the intrinsic time scales of the system under analysis.

### 3.3 Forbidden and missing ordinal patterns

As shown recently by Amigó et al. [7,24,32,33], in the case of deterministic chaotic one-dimensional maps, not all the possible ordinal patterns can be effectively materialized into orbits, which in a sense makes these patterns "forbidden". The dynamics of the map makes them never occurring. More precisely, for chaotic map time series a minimum pattern length $D_{min}$ is necessary to be considered in order to observe forbidden patterns [34]. The $D_{min}$ is characteristic for the chaotic maps under analysis. One expects in general that higher-dimensional chaotic dynamical systems (maps) will also exhibit forbidden patterns. Indeed, the existence of these *forbidden ordinal patterns* becomes a persistent fact that can be regarded as a "new" dynamical property. Thus, for a fixed pattern-length (embedding dimension $D \geq D_{min}$) the number of forbidden patterns of a time series (unobserved patterns) is independent of the series' length $N$. Note that this independence is not shared by other properties of the series, such as proximity and correlation, which die out with time [7,24].

Stochastic processes could also display forbidden patterns [10,35]. However, in the case of either *uncorrelated* (white noise) or *correlated stochastic processes* (noise with power spectrum $f^{-k}$ with $k > 0$; fractional Brownian motion and fractional Gaussian noise) it can be numerically ascertained that *no* forbidden patterns emerge. For time series generated by *unconstrained stochastic processes* (uncorrelated processes) every ordinal pattern has the same probability of appearance [7,24,32,33]. Indeed, if the data set is long enough, all ordinal patterns will eventually appear. In this case, as the number of time series observations increases, the associated PDF becomes uniform, and the number of observed patterns will depend only on the time series length $N$. The existence of a non-observed ordinal pattern does not qualify it as *"forbidden"*, only as *"missing"*, and this could be due to the time series finite length.

For correlated stochastic processes, the probability of observing a specific individual pattern depends not only on the time series' length $N$, but also on the correlation structure [36]. For highly correlated series, patterns which belong to rapid changes (volatile behaviour) will be much less frequent than monotonous patterns.

Measured data from real systems always possess a stochastic component due to the omnipresence of dynamical noise [37–39]. Thus, the existence of "missing ordinal patterns" could be either related to stochastic processes (correlated or uncorrelated) or to deterministic noisy processes. Observational time series will most likely be a mixture of both.

## 4 Results

We now show results for 498 long runoff series together with corresponding ones for $k$-noise. We mainly show results for $D = 6$, as a compromise to include as much dynamics as possible and the finite data set length. However, we varied $D$ between $D = 3$ and $D = 7$. There is no point to try $D = 2$ since in this case, lower and upper limit curves are identical, implying that the relation between entropy and complexity is a deterministic function. For $D > 7$, all runoff time series are too short, i.e. $N < D!$. The issue of finite size is already relevant at $D = 7$. For $D = 6$, it holds in all cases that $N \gg D!$.

We also fix $\tau = 1$ (one day). The choice of the temporal lag also impacts the results, as more and more of the correlation structure is recognized by the ordinal patterns with increasing $\tau$. There is also a difference between decimating the series or aggregating it, e.g. to weekly or monthly values. These methodological considerations are, however, outside the scope of the paper.

Having chosen these two parameters, the calculation of the order statistics and then of permutation entropy and complexity is straightforward. An overview of the locations of our runoff series in the Complexity-Entropy-Causality Plane (CECP) is presented in Fig. 2. In this figure, the upper and lower limit curves for the complexity are also included. The result for the $k$-noise (blue) is obtained as the average of 100 independent realizations at each fixed $k$ (from 0 to 5 in steps of $\Delta k = 0.1$) and a length of $N = 10^5$; with this setup, the $(\mathcal{H}, \mathcal{C})$ values for the $k$ noise happen to fall onto a single simple curve. For $k = 0$, the uncorrelated white noise case, the result is in the lower right corner $(\mathcal{H} = 1, \mathcal{C} = 0)$. For increasing values of $k$, $\mathcal{H}$ is decreasing, whereas $C$ exhibits a maximum. This curve represents a parameter-free asymptotic characterization of $k$-noise behavior in the CECP.

Values for the runoff series are at intermediate complexity for all $D$ investigated. Values close to the upper limit curve are absent, which distinguishes runoff data from deterministic chaotic maps [9]. For the majority of rivers, the position in the CECP is above the curve defined by the $k$-noise; this, however depends on the choice of $D$: starting with lower values than $k$-noise for $D = 3$, with increasing $D$, the $C$ values

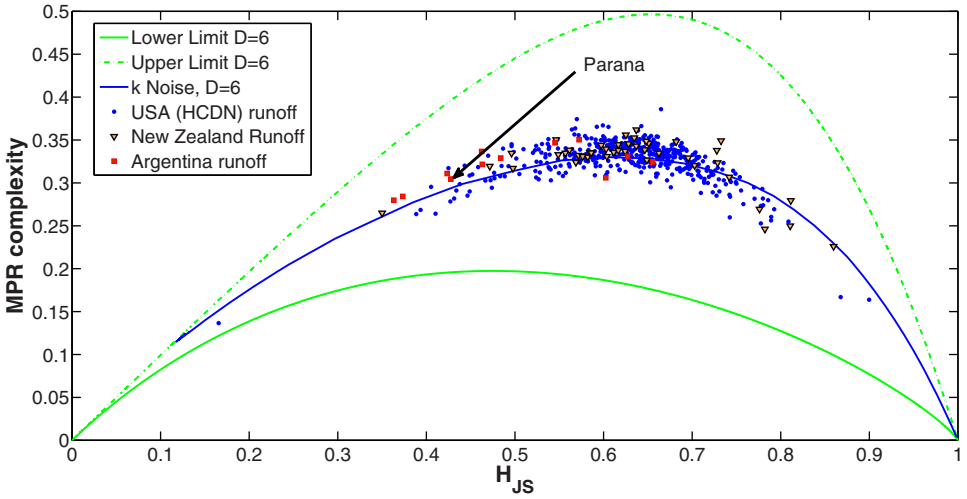


**Fig. 2.** The CECP at $D = 6$. The three curves show the upper and lower limit of the complexity at a given entropy for any process, and the results of a large number of $k$-noise simulations (in blue). Here, $k$ runs from 0 (lower right corner) to 5 with step of $\Delta k = 0.1$. Points are individual runoff time series.

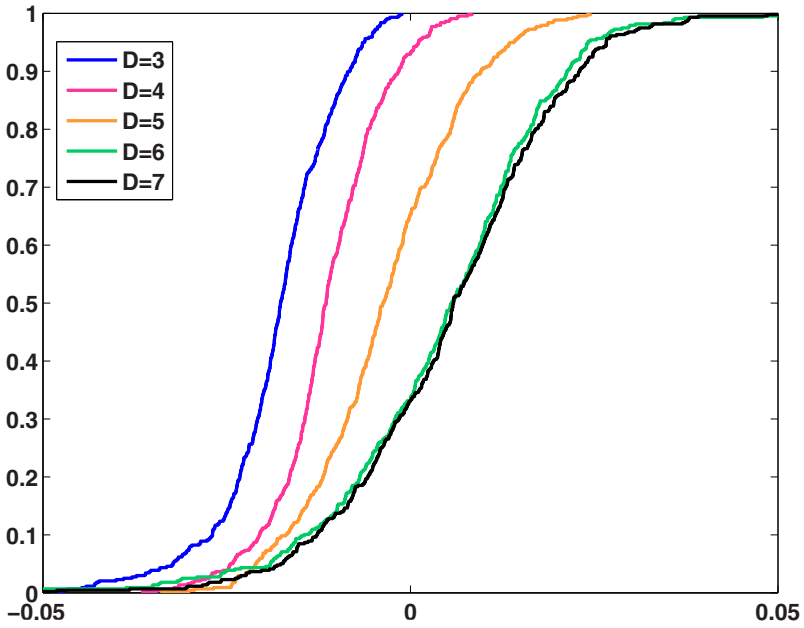


**Fig. 3.** Cumulative distributions of the differences between complexity of runoff and $k$ noise with identical entropy for different embedding dimensions.

of runoff series are increasing faster than those of $k$-noise. We quantified this observation by calculating the complexity along the $k$-noise curve at each entropy value of the runoff series, using shape-preserving interpolation. We then took the difference between the complexity obtained from each runoff series and the $k$-noise. Cumulative distributions of these differences for $D = 3 - 7$ are shown in Fig. 3. With increasing $D$, this distribution shifts to the right and cross the zero, implying that runoff is

more complex than $k$-noise. The difference between the curves is minor for $D = 6$ and $D = 7$, demonstrating the onset of finite size effects.

It is apparent that the autocorrelation structure of runoff series is more complex than that of $k$-noise (which is analytically known), a fact which is however unimportant at very small time lags. At $D = 6$, $k$-noise still seems to be sufficient as a rough characterization of runoff data in the CECP, as the latter by and large seem to be scattered around that curve. As we cannot reach large time lags (use large $D$ values) due to insufficient time series length, we are interested in how good exactly is the assumption that runoff data are simply $k$-noise realizations at small lags?

We use one specific time series, the runoff of the Paraná river in Argentina (gauge located at $31°51'41''S$, $60°30'15''W$), for a detailed comparison between observations and $k$-noise. It has a length of $N_p = 37712$ data values at daily resolution, or more than 103 years (spanning from 1904 to 2007). For $D = 6$, the time series has $\mathcal{H} = 0.4277$, $\mathcal{C} = 0.3045$, as indicated in Fig. 2. Large catchments such as this one typically have smaller values for $\mathcal{H}$ – the signal is less noisy due to integration over larger time scales. An extreme example is the Amazon Basin, the largest catchment worldwide, which is the outlier to the far left of the CECP. The complexity values are to a lesser extent related to the catchment size. The one for the Paraná river is higher than the corresponding $k$-noise one, but what does this difference imply?

There are several strategies to investigate to which degree runoff time series resemble $k$-noise or not. One obvious way is to compare the time series graphs. Before that, one has to determine or estimate the value of $k$ which comes closest to the observed series. This can be done in at least one of three ways: (1) to find a $k$ which minimizes the Euclidean distance between the blue curve in Fig. 2 and the observed data point for Paraná; (2) to find the value of $k$ where the number of missing patterns is the same as for the Paraná series; (3) to determine the decrease of missing patterns as a function of increasing data length up to $N_p$, fit this decrease to an exponential function, and determine the $k$ value where the slope is the same as that of the observed series.

The first way is straightforward after interpolating the $k$-noise curve using shape-preserving Hermitean polynomials [40]. We then use a Levenberg-Marquardt algorithm to minimize the distance between the point for the Paraná river and the $k$-noise curve, and read off the exponent from the Hermitean interpolation. The result is $k_{est,1} = 3.117$, indicating rather strong and long-ranged correlations in the runoff record as expected – note that the lag-1 autocorrelation of such a noise is 1 within one part in $10^5$.

For the second method, we calculate the number of missing patterns for a large number of $k$-noise realizations (1000 for each chosen $k$ value), keeping the length fixed to $N_p$ each time. The result at fixed $k$ is surprisingly robust (Fig. 4). In addition to the mean number of patterns missing, the extremes from these simulations are also shown. For $k = 2$, almost no patterns are missing, or all $D! = 720$ patterns are occurring. For white noise, $k = 0$, all of these have the same probability in addition. There is a sharp and relatively well-defined increase until the number of missing patterns saturates to an asymptotic value of 694, apparently independent of $k$. However, this is just an effect of the finite length of each realization. For increasing data set length, all but one of the remaining 26 patterns eventually disappear as well [33], the higher the $k$ value the slower. The number of missing patterns for the Paraná series is $N_{miss} = 212$. From an interpolation of the mean curve of Fig. 4, we estimate $k_{est,2} = 3.273$.

The two exponent estimates are not drastically different; for convenience, we continue with $k_{est,2}$ in the following. The entropy and complexity values for $k_{est,2}$-noise are $\mathcal{H}(k_{est,2}) = 0.3447$ and $\mathcal{C}(k_{est,2}) = 0.2582$. These values are 81% and 85% of the corresponding ones for the Paraná series. Where does this difference come from?

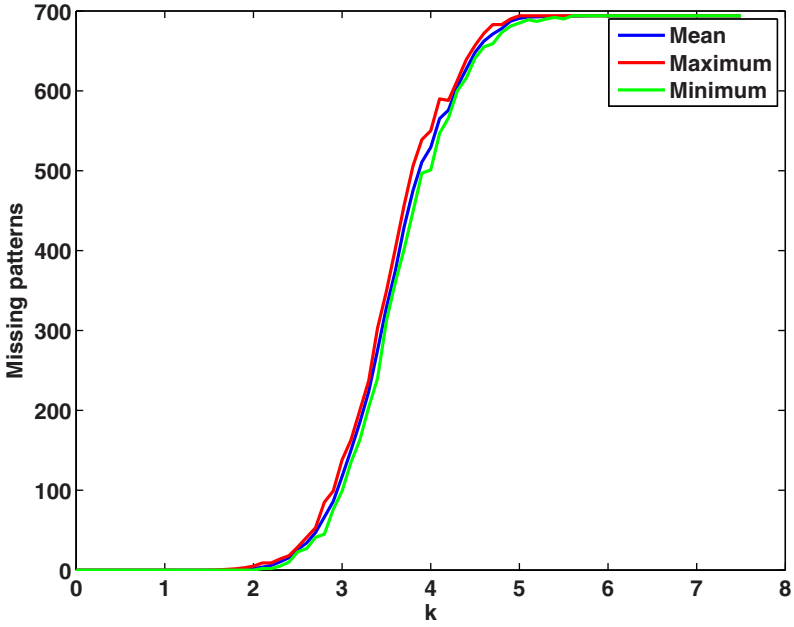


**Fig. 4.** The number of missing patterns in $k$-noise. Mean, maximum and minimum values based on 1000 realizations for each $k$ value. The Bandt-Pompe parameters were: $D = 6$, $\tau = 1$ and time series length $N = 37712$.

The third way is maybe the most intrigued one. At a chosen value of $k$, we determine the number of missing patterns in subseries of given length $L \leq N_p$, and average them. We increase $L$ from 1000 to $N_p$ in small steps ($\Delta L = 100$). Then, as suggested by [10,24,33], we fit the missing patterns ($mp$) obtained to

$$mp(L) = Ae^{R \cdot L} \tag{6}$$

and get the slopes $R$ for different $k$ values using least squares. The result is shown in Fig. 5. For the Paraná series, we obtain a slope $R = -2.46 \cdot 10^{-5}$. Inverting the curve of Fig. 5 leads to a $k_{est,3} = 3.2936$. Complexity and entropy for $k_{est,3}$ noise are 77% and 84% of that of the Paraná series.

Unfortunately, the simple exponential in Eq. (6), although intuitive, does not represent the behaviour of the $mp(L)$ very well. This is demonstrated in Fig. 6. The two exponential fits are almost parallel-shifted versions of each other since the $k_{est,3}$ was tuned to reproduce the slope. Although the $r^2$ values are high, it is obvious that a simple two-parameter fit is not sufficient to reproduce the calculated missing patterns of either of them; the exponential fit seems to overestimate the asymptotic decline in patterns. The two curves also exhibit slightly different behaviour in particular for large data length.

Visual inspection of the observed runoff and a rescaled version of one realization of $k_{est,2}$-noise reveals that the two are of very different nature (Fig. 7). High-frequency fluctuations are dominating the runoff time series, whereas the $k$-noise is smoother. The two series are completely unrelated, i.e. no fitting has been performed, and every realization of a $k$-noise process looks differently. However, the general appearance, in particular the lack of strong high frequency fluctuations, is a common feature for such high $k$ values as used here.

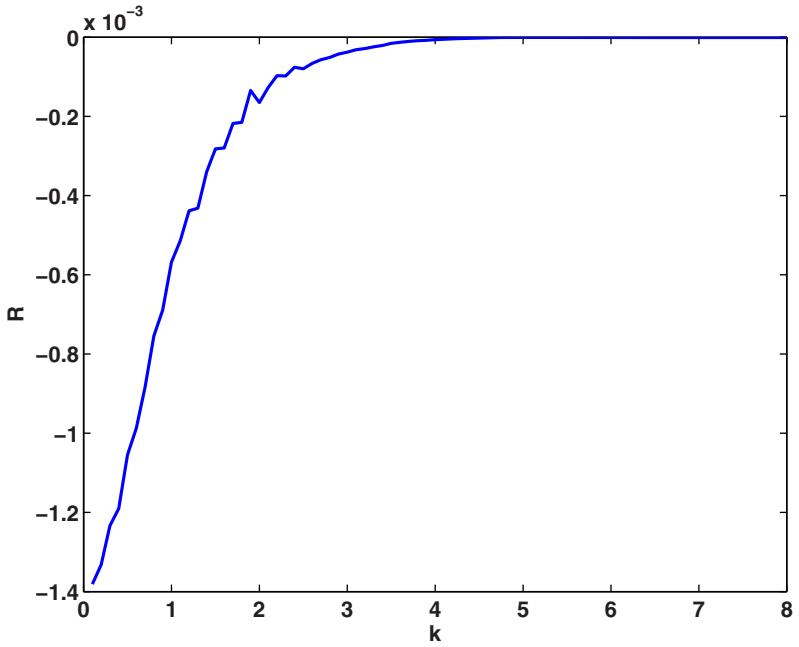


**Fig. 5.** The slope of the relationship between the number of missing patterns and $k$, using data set lengths between 1000 and 37712. The exponential in Eq. (6) was used to obtain least-squares estimates.

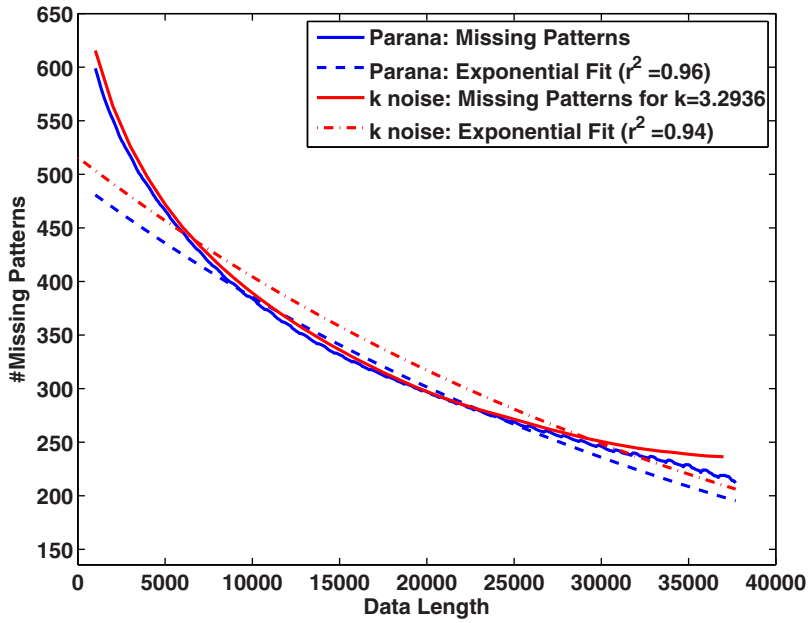


**Fig. 6.** The decay of missing patterns as a function of increasing time series length. Dotted lines are the result of least squares fits of the respective curves to an exponential model.

Another, more traditional possibility to compare runoff time series and $k$-noise would have been using the autocorrelation function. However, we are inspecting the observations and the noise process from the perspective of statistical complexity and

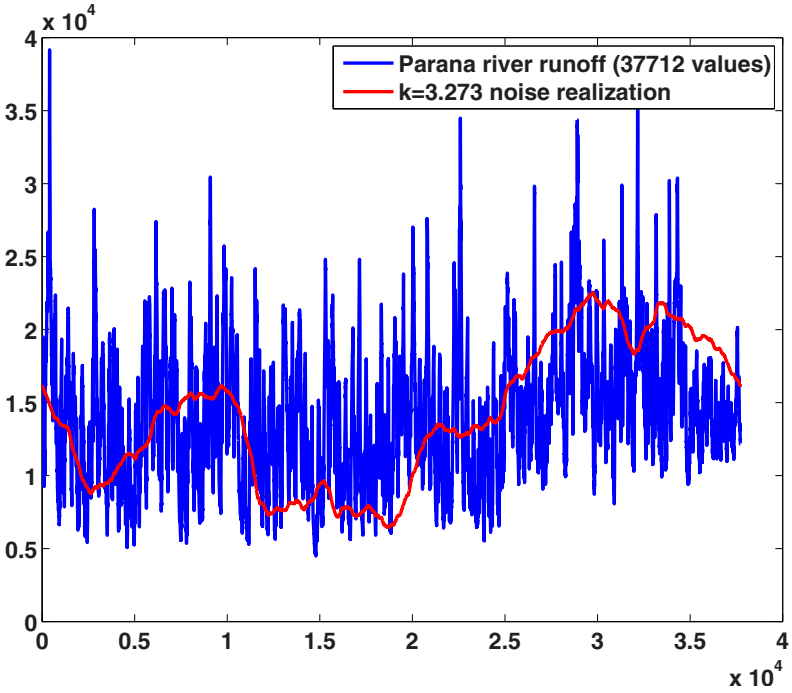


**Fig. 7.** Comparison of the Paraná runoff time series and one realization of the $k$ noise process where $k = k_{est,2}$. Mean value and standard deviation have been scaled to the same values for both series. Note that no fitting of any kind has been performed.

order patterns, since the aim is to conclude on deterministic parts of the observed series. Thus, we investigate the respective order patterns in more detail now. And since $k_{est,2}$-noise reproduces the amount of missing patterns (212) at length $N_p$, we continue with that value in the following.

The patterns are numbered with an index $n$ (running from 1 to 720) in a unique manner, where we adopt the convention of Keller based on inversion numbers [27]. For Paraná, the number of patterns are strongly inversely related to the number of inversions (changes in the direction of increase and decrease) within a pattern, cf. Fig. 8. Here, the total counts within the Paraná series are displayed for each possible number of inversions. Note that there are only two patterns with 0 inversions (no. 1 and 720 in Keller's convention). The number of patterns with a given number of inversions for $D = 6$ are 2, 60, 236, 300 and 122 for 0, 1, 2, 3 and 4 (the maximum possible) inversions. Using $k$-noise realizations with the same number of missing patterns as for Paraná (i.e., working with $k_{est,2}$), we can investigate and compare the pattern distributions. The distributions are strikingly uneven, with a few huge peaks and a bunch of very small values.

The highest counts by far refer to pattern no. 1 and no. 720; the first one is the strictly increasing patterns ($x_1 < x_2 < \cdots < x_6$ or {123456} for short) whereas the second one is strictly decreasing {654321}. The latter one alone occurs in 37 % of all cases for the runoff. This reflects the strong autocorrelation of the series, where rapid changes in the sign of differences are unlikely.

Closer inspection of the distribution peaks of the $k$-noise reveals that they occur in pairs with almost identical peak heights. Examples for such pairs are patterns no. 361 – sequence {123465} and no. 719 – sequence {564321}; or no. 3 – sequence {312456}

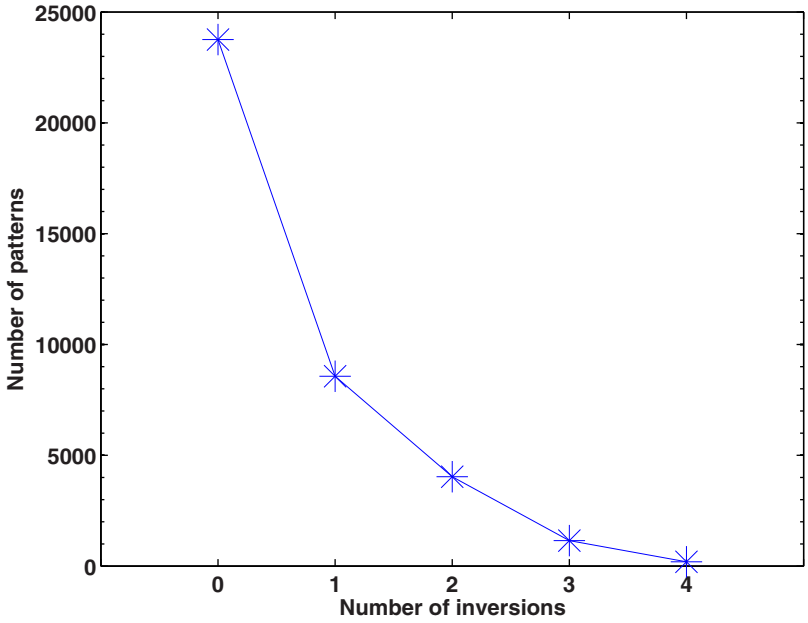


**Fig. 8.** The number of patterns occurring in the Paraná runoff series as a function of the number of inversions within each pattern.

and no. 240 – sequence {654213}. The common property of these pairs of patterns is that their members are *time reversals* of each other: one is the other read backwards.

We generate $k$-noise data using a Gaussian random number generator and distort its power spectrum before transforming back. This transformation is phase-blind and the generated data are thus time-reversal invariant: rising limbs appear with the same probability and temporal dynamics as falling limbs. The small differences between the pairs are thus just finite size fluctuations which should vanish for very long series.

Runoff data, on the other hand, are decidedly temporally asymmetric. Fast rises during or after heavy rainfall are followed by slow recessions. "Fast" and "slow" depend on the watershed area, ground water volume, and other factors, but the asymmetry is always present.

We further quantify this difference between $k$-noise and runoff by calculating and visualizing an index reflecting the relative difference between all pairs of time reversal patterns. Since very rare patterns are supposedly just due to finite size effects, we restrict the analysis to pattern pairs where the less frequent one has at least 10 counts in the records.

We first define the time reversed version of a time series $x$ of length $N$.

$$x_{rev}(i) = x(N+1-i) \ , \qquad i = 1, 2, \cdots, N \ . \tag{7}$$

For each pattern $p(r)$, where $r$ is denoting the pattern number ($r = 1, \cdots, 720$ in our case), we determine its temporal reverse, $p_{rev}(r)$ in the Keller coding scheme. Then, we calculate the relative asymmetry between the time-forward and the time-backward version of each pattern in the time series as

$$A(r) = \frac{|n\{p(r)\} - n\{p_{rev}(r)\}|}{n\{p(r)\} + n\{p_{rev}(r)\}} \tag{8}$$

where $n\{p(r)\}$ denotes the counts for the pattern $p(r)$.

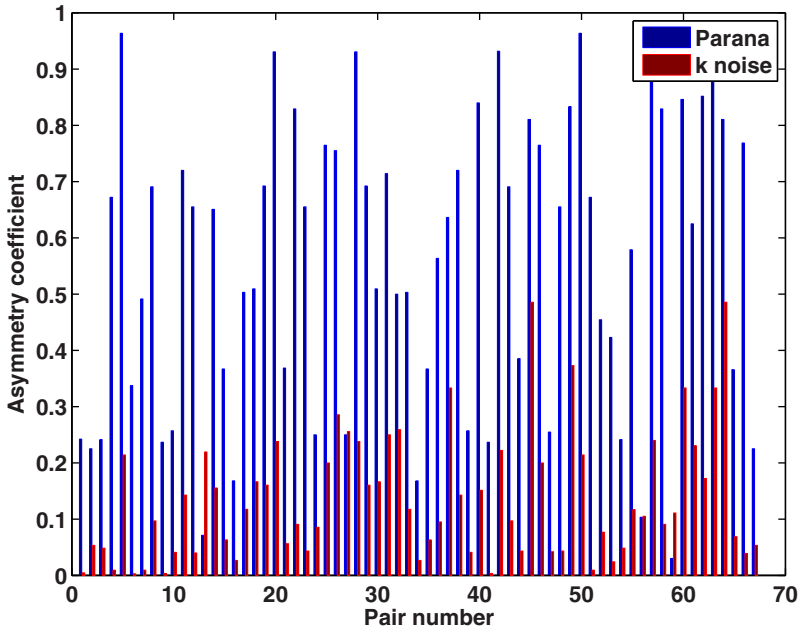


**Fig. 9.** Values of the time-reversal asymmetry coefficient for all pairs with at least 10 counts in every entry.

Comparing two time series from the perspective of their temporal asymmetry, in general the number and type of patterns occurring in Eq. (8) will be different, due to the threshold of at least 10 counts in both patterns. For the case at hand, there are 107 patterns fulfilling this criterion for noise with $k = k_{est,2}$, and 79 patterns for the Paraná river. The intersection between these two sets of patterns leads to 67 common patterns where both $A(r)$ values can be calculated. These values are shown in Fig. 9.

The asymmetry of the runoff data is very obvious and higher than that of the $k$ noise in 63 out of 67 cases. We can also calculate an average asymmetry summing the expression in Eq. (8) and dividing by 67: $< A_{runoff} >= 0.6638$ and $< A_{k-noise} >= 0.1427$. The latter value indicates a benchmark or significance threshold, since $k$-noise is temporally symmetric. One might develop a numerical significance test for temporal asymmetry based on that.

How do the entropy and complexity values react to time reversal in both cases? Table 1 provides values for the time-forward and the time-backward version of each series. Both entropy and complexity of the runoff increase through the operation of time reversal. For the $k$-noise, there are also changes which are, however, an order or magnitude smaller and in both directions. The asymmetry in time for the runoff data is a system property, absent in the $k$ noise data, and thus a feature distinguishing hydrological data from purely stochastic processes.

The results obtained from the specific case also hold for the whole set of records. We calculated the mean asymmetry coefficient for all 498 runoff time series and for the same number of $k$-noise series with length $N = 10000$ and exponent $k = 3$, a typical value for the runoff records. The histogram of the $< A >$ distributions are shown in Fig. 10. The distribution for runoff is much broader, and there is virtually no overlap between the two histograms; they are completely different. Thus, pronounced temporal asymmetry is a pertinent feature of the observed records.

**Table 1.** Entropy and complexity for original, time-reversed and symmetrized versions of the series.

| *Series* | *Paraná* | $k$-noise |
|---|---|---|
| $\mathcal{H}$ forward | 0.4277 | 0.3583 |
| $\mathcal{H}$ backward | 0.4418 | 0.3586 |
| Difference, % | **3.3** | **0.09** |
| $\mathcal{C}$ forward | 0.3045 | 0.2656 |
| $\mathcal{C}$ backward | 0.3085 | 0.2651 |
| Difference, % | **1.3** | **−0.16** |
| $\mathcal{H}$ symmetrized | 0.4362 | 0.3592 |
| Difference, % | **1.99** | **0.25** |
| $\mathcal{C}$ symmetrized | 0.3058 | 0.2648 |
| Difference, % | **0.43** | **−0.30** |

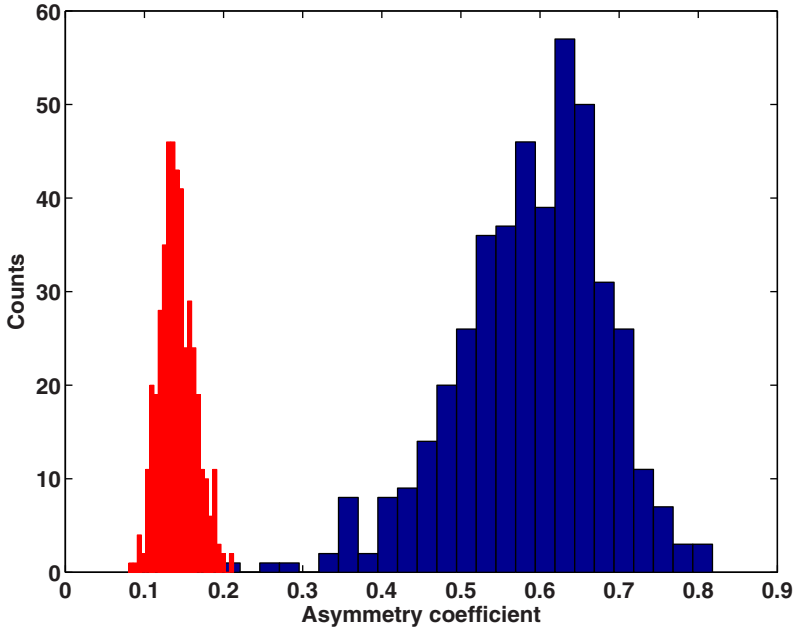


**Fig. 10.** Histograms of the mean asymmetry coefficients for the runoff time series (blue) and an identical number of $k$ noise realizations with $k = 3$ and length 10000 data points (red).

This temporal asymmetry can be eliminated through symmetrizing the data: stitching together original and time-reversed time series

$$x_{sym} = x \circ x_{rev}, \tag{9}$$

where $\circ$ denotes concatenation, leads to explicitly time-symmetric time series. Their $\mathcal{H}$ and $\mathcal{C}$ values are reported in Table 1 as well. For the runoff, they are intermediate between forward and backward version; for $k$-noise, no clear picture emerges, and the changes are once again rather small.

As expected, the number of missing patterns is much lower in the symmetrized versions; if one pattern has been present in the forward version, its reversed version is present as well now in the symmetrized version. Remembering that the $k_{est,2}$ was tuned to reproduce the number of missing patterns in the runoff data (212), it comes

at no surprise that after symmetrization, the number of missing patterns is nearly identical now (runoff: 132, $k$-noise: 131). The pattern missing in the symmetrized runoff have 2 inversions (2 patterns), 3 inversions (68), or 4 inversions (62), thus they are suppressed due to the strong correlations in the series. However, only 62 of the missing patterns in the runoff are also missing in the $k$-noise. The remaining patterns which do not occur in the runoff but do occur in the $k$-noise have counts in the range from 1 to 15, with an average of only 2.3. This could easily be non-significant spurious fluctuations. Still, the $\mathcal{H}$ ($\mathcal{C}$) of runoff is 121 (115) % of that of the $k$-noise in the symmetrized versions (Table 1). This indicates that there are further differences between runoff data and $k$-noise unrelated to temporal asymmetry.

## 5 Discussion and outlook

We have characterized river runoff data in terms of permutation entropy and statistical complexity. At daily resolution, they exhibit intermediate to high entropy and a complexity between correlated stochastic processes and deterministic chaotic series. This is an indication that these data are a mixture of random and deterministic parts. We compared the observations with a reference stochastic process, the $k$-noise, which appears as a particularly simple structure (a one-dimensional curve) in the CECP. Even when tuned to the closest exponent $k$, runoff data show qualitatively different behaviour compared to the stochastic process. This is most clearly reflected in the temporal asymmetry, for which we developed a quantitative index based on order patterns.

Even when symmetrized, a difference between runoff and stochastic process remains, with runoff data typically at higher complexity values. Work in progress should reveal the origin and interpretation of this difference. We will investigate the equivalence class of time series which have identical entropy and complexity value. To that end, we are currently developing a generator for time series at any desired location in the CECP and analyze their properties, e.g. their correlational structure.

The framework is also a very suitable approach for model-data comparisons. Short-term correlated time series models such as AR(p) are generally not expected to be good candidates for simulating runoff on longer time scales, although they are used for short-term prediction. This analysis shows that they fail qualitatively also on short time scales. For more sophisticated, process-based models, the CECP comparison provides a stringent challenge.

Another important application domain for nonlinear analysis methods is the *classification* of catchments or hydrological systems in general. Here, an approach based on the correlation dimension revealed regional differences in the western U.S. [2]. Information and complexity measures based on parametric partitioning showed differences potentially related to geological and vegetation history [41]. We continue our search for determinism in runoff time series from a CECP perspective, and try further to extract deterministic properties from the records.

We would like to thank J. Kurths for the invitation to join this EPJ Special Topics issue on Order Statistics. We are grateful to A. Pasquini for providing runoff data from Argentina. We appreciate the support of T. Davie, B. Fahey, R. Jackson, A. McKerchar and R. Woods who provided runoff data from experimental catchments and hydrological monitoring in New Zealand. H. Lange thanks F. Serinaldi for earlier discussions on hydrology and complexity. O. A. Rosso gratefully acknowledges support from CONICET, Argentina and CNPq fellowship, Brazil.

## References


1. J.J. McDonnell, R. Woods, J. Hydrology **299**, 2 (2003)
2. B. Sivakumar, V.P. Singh, Hydro. Earth Syst. Sci. **16**, 4119 (2012)
3. T. Wagener, M. Sivapalan, P.P. Troch, R. Woods, Geography Compass **1**, 901 (2007)
4. R. Woods, Adv. Water Resources **26**, 295 (2003)
5. N.S. Davies, M.R. Gibling, Nature Geosci. **4**, 629 (2011)
6. C. Bandt, B. Pompe, Phys. Rev. Lett. **88**, 174102 (2002)
7. J.M. Amigó, *Permutation complexity in dynamical systems* (Springer-Verlag, Berlin, 2010)
8. M.T. Martín, A. Plastino, O.A. Rosso, Physica A **369**, 439 (2006)
9. O.A. Rosso, H.A. Larrondo, M.T. Martín, A. Plastino, M.A. Fuentes, Phys. Rev. Lett. **99**, 154102 (2007)
10. O.A. Rosso, L.C. Carpi, P.M. Saco, M. Gómez Ravetti, A. Plastino, H.A. Larrondo, Physica A **391**, 42 (2012)
11. H.D.I. Abarbanel, *Analysis of Observed Chaotic Data* (Springer-Verlag, New York, 1996)
12. A.N. Kolmogorov, Dokl. Akad. Nauk. SSSR **119**, 861 (1958)
13. Y.G. Sinai, Dokl. Akad. Nauk. SSSR **124**, 768 (1959)
14. M. Zanin, L. Zunino, O.A. Rosso, D. Papo, Entropy **14**, 1553 (2012)
15. C.E. Shannon, Bell Syst. Technol. J. **27**, 379 (1948)
16. D.P. Feldman, J.P. Crutchfield, Phys. Lett. A **238**, 244 (1998)
17. D.P. Feldman, C.S. McTague, J.P. Crutchfield, Chaos **18**, 043106 (2008)
18. P.W. Lamberti, M.T. Martín, A. Plastino, O.A. Rosso, Physica A **334**, 119 (2004)
19. R. López-Ruiz, H.L. Mancini, X. Calbet, Phys. Lett. A **209**, 321 (1995)
20. I. Grosse, P. Bernaola-Galván, P. Carpena, R. Román-Roldán, J. Oliver, H.E. Stanley, Phys. Rev. E **65**, 041905 (2002)
21. A.R. Plastino, A. Plastino, Phys. Rev. E **54**, 4423 (1996)
22. O.A. Rosso C. Masoller, Phys. Rev. E **79**, 040106(R) (2009)
23. O.A. Rosso, C. Masoller, Eur. Phys. J. B **69**, 37 (2009)
24. J.M. Amigó, S. Zambrano M.A.F. Sanjuán, Europhys. Lett. **79**, 50001 (2007)
25. P.M. Saco, L.C. Carpi, A. Figliola, E. Serrano, O.A. Rosso, Physica A **389**, 5022 (2010)
26. K. Keller, M. Sinn, Physica A **356**, 114 (2005)
27. K. Keller, M. Sinn, J. Emonds, Stoch. Dyn. **7**, 247 (2007)
28. M.C. Soriano, L. Zunino, L. Larger, I. Fischer, C.R. Mirasso, Opt. Lett. **36**, 2212 (2011)
29. M.C. Soriano, L. Zunino, O.A. Rosso, I. Fischer, C.R. Mirasso, IEEE J. Quantum Electron. **47**, 252 (2011)
30. L. Zunino, M.C. Soriano, I. Fischer, O.A. Rosso, C.R. Mirasso, Phys. Rev. E **82**, 046212 (2010)
31. L. Zunino, M.C. Soriano, O.A. Rosso, Phys. Rev. E **86**, 046210 (2012)
32. J.M. Amigó, L. Kocarev, J. Szczepanski, Phys. Lett. A **355**, 27 (2006)
33. J.M. Amigó, S. Zambrano, M.A.F. Sanjuán, Europhys. Lett. **83**, 60005 (2008)
34. O.A. Rosso, F. Olivares, L. Zunino, L. De Micco, A.L.L. Aquino, A. Plastino, A. Larrondo, Eur. Phys. J. B **86**, 116 (2013)
35. O.A. Rosso, L.C. Carpi, P.M. Saco, M. Gómez Ravetti, H.A. Larrondo, A. Plastino, Eur. Phys. J. B **85**, 419 (2012)
36. L.C. Carpi, P.M. Saco, O.A. Rosso, Physica A **389**, 2020 (2010)
37. H. Wold, *A Study in the Analysis of Stationary Time Series* (Almqvist and Wiksell, Upsala, Sweden, 1938)
38. J. Kurths, H. Herzel, Physica D **25**, 165 (1987)
39. S. Cambanis, C.D. Hardin, A. Weron, Probab. Theory Relat. Fields **79**, 1 (1988)
40. F.N. Fritsch, R.E. Carlson, SIAM J. Numer. Anal. **17**, 238 (1980)
41. M. Hauhs, H. Lange, Geogr. Compass **2**, 235 (2008)