Technical note: c-u-curve: A method to analyse, classify and compare dynamical systems by uncertainty and complexity

Uwe Ehret¹, Pankaj Dey²

¹Institute of Water Resources and River Basin Management, Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany
²Interdisciplinary Centre for Water Research, Indian Institute of Science, Bangalore, India

Correspondence to: Uwe Ehret (uwe.ehret@kit.edu)

Abstract. We propose and provide a proof-of-concept of a method to analyse, classify and compare dynamical systems of arbitrary dimension by the two key features uncertainty and complexity. It starts by subdividing the system’s time-trajectory into a number of time slices. For all values in a time slice, the Shannon information entropy is calculated, measuring within-slice variability. System uncertainty is then expressed by the mean entropy of all time slices. We define system complexity as “uncertainty about uncertainty”, and express it by the entropy of the entropies of all time slices. Calculating and plotting uncertainty u and complexity c for many different numbers of time slices yields the c-u-curve. Systems can be analysed, compared and classified by the c-u-curve in terms of i) its overall shape, ii) mean and maximum uncertainty, iii) mean and maximum complexity, and iv) its characteristic time scale expressed by the width of the time slice for which maximum complexity occurs. We demonstrate the method at the example of both synthetic and real-world time series (constant, random noise, Lorenz attractor, precipitation and streamflow) and show that the shape and properties of the respective c-u-curves clearly reflect the particular characteristics of each time series. For the hydrological time series we also show that the c-u-curve characteristics are in accordance with hydrological system understanding. We conclude that the c-u-curve method can be used to analyse, classify and compare dynamical systems. In particular, it can be used to classify hydrological systems into similar groups, a precondition for regionalization, and it can be used as a diagnostic measure which can be used as an objective function in hydrological model calibration. Distinctive features of the method are i) that it is based on unit-free probabilities, thus permitting application to any kind of data, ii) that it is bounded, iii) that it naturally expands from single- to multivariate systems, and iv) that it is applicable to both deterministic and probabilistic value representations, permitting e.g. application to ensemble model predictions.

Keywords
Complexity, uncertainty, entropy, dynamical system analysis, catchment classification, catchment similarity.

1 Introduction

In the earth sciences, many systems of interest are dynamical, i.e. their states are ordered by time and evolve as a function of time. The theory of dynamical systems (Forrester, 1968; Strogatz, 1994) therefore has proven useful across a wide range of earth science systems and problems such as weather prediction (Lorenz, 1969), ecology (Hastings et al., 1993; Bossel, 1986), hydrology (Koutsoyiannis, 2006), geomorphology (Phillips, 2006) and coupled human-ecological systems (Bossel, 2007).

Key characteristics of dynamical systems include their mean states (e.g. climatic mean values in the atmospheric sciences), their variability (e.g. annual minimum and maximum streamflow in hydrology) and their complexity (e.g. population dynamics in ecological predator-prey cycles). Interestingly, despite its importance and widespread use there is to date no single agreed-upon definition and interpretation of complexity (Prokopenko et al., 2009; Gell-Mann, 1995), and many different concepts are employed across disciplines.

Characterizing dynamical systems by few and meaningful statistics representing the above-mentioned key features is important for several reasons: System classification, intercomparison and similarity analysis is a precondition for the transfer
of knowledge from well-known to poorly-known systems or situations (see e.g. Wagener et al., 2007; Sawicz et al., 2011; and Seibert et al., 2017 for applications in hydrology). Further, dynamical system analysis helps detecting and quantifying nonstationarity, a key aspect in the context of global change (Ehret et al., 2014), and it is important for evaluating the realism of dynamical system models and for guiding their targeted improvement (Moriasi et al., 2007; Yapo et al., 1998).

In this paper, we address the task of parsimonious yet comprehensive characterization of dynamical systems by proposing a method based on concepts of information theory. It comprises both variability and complexity, and adopts the view that the complexity of a time series is the overall variability of its variabilities in sub-periods. We use examples from hydrology, as due to the multitude of subsystems and processes involved, most hydrological systems classify as complex systems (Dooge, 1986). Hydrological systems have been analysed and classified in terms of their complexity by Jenerette et al. (2012), Jovanovic et al. (2017), Ossola et al. (2015), Bras (2015), Engelhardt et al. (2009), Pande and Moayeri (2018), Sivakumar and Singh (2012) and Sivakumar et al. (2007), among others. In the same context, concepts from information theory have been applied by Pachepsky et al. (2006), Hauhs and Lange (2008), Zhou et al., (2012), Castillo et al., (2015), and recently by Dey and Mujumdar (2021). Information-based approaches rely on log-transformed probability distributions of the quantities of interest, and are thus independent of the units of the data. Compared to methods relying directly on the data values, this poses an advantage in terms of generality and comparability across disciplines. Being rooted in information theory, the method we propose in this paper makes use of this advantage. Compared to existing information-based methods, it offers the additional advantages of jointly considering two key characteristics of dynamical systems, uncertainty and complexity, and of being applicable to multivariate and ensemble data sets.

The remainder of the text is organized as follows: In Sect. 2, we present all steps of the method. In Sect. 3, we apply the method to both synthetic time series and observed hydrological data. We use these examples to describe the properties, interpretations and potential uses of the c-u-curve method. We summarize the method, discuss its limitations and draw conclusions in the final Sect. 4.

2 Method

Please note that in what follows, for clarity we introduce the method at the example of univariate time series with deterministic values, and we also use the corresponding variable notation. In Ehret (2022) we also provide generalized application examples for multivariate and ensemble cases, and the related generalized code. Also, we calculate discrete entropy based on a uniform binning approach. We do so as it has some useful properties (ease of interpretation is one of them) compared to calculating continuous entropy. Nevertheless, the method can also be used with non-uniform binning or continuous representations of data-distributions. For a detailed discussion of discrete vs. continuous entropy, see Azmi et al. (2021) and references therein.

2.1 Method description

Applying the method to a given time series with overall nt time steps consists of a number of steps and related choices: At first, for each variate involved a suitable discretization (binning) scheme is chosen. The bins must cover the entire value range, and their total number nub can be chosen according to a user’s demands on data-resolution. Next, the time series is divided into a number of ns time slices. The slices must be mutually exclusive and together must cover the time series. The slices are preferably, but not necessarily, of uniform width. Next, separately for each slice, a discrete probability distribution (histogram) is calculated using the data in the slice and the chosen binning scheme. From the so-obtained histogram, the Shannon information entropy $H$ (Shannon, 1948) is calculated following Eq. (1),

$$H(X) = - \sum_{v=1}^{nub} p(x_{vb}) \cdot \log_2(p(x_{vb}))$$  (1)
where \( p(x_{vb}) \) refers to the probability of variate value \( x \) falling into bin \( vb \), and \( n_{vb} \) is the total number of value bins. Entropy measures data variability or uncertainty in bit, with the intuitive interpretation as “the minimum number of binary (Yes/No) questions needed to be asked to correctly guess values drawn from a known distribution”. Cover and Thomas (2006) provide an excellent introduction to information theory, applications to hydrology and hydrometeorology are e.g. presented in Singh (2013) and Neuper and Ehret (2019). As entropy values may differ between slices, an overall uncertainty estimate for all slices is calculated as the expected value of all slice entropies. For equal-width slices, this is mean entropy according to Eq. (2),

\[
\text{Uncertainty} = E(H(X)) = \bar{H}(X) = \frac{1}{ns} \cdot \sum_{s=1}^{ns} H_s(X)
\]

where \( s \) refers to a particular slice of all \( ns \) time slices. The so-defined uncertainty measures average within-slice variability of the data, i.e. uncertainty of the time series as seen through the lens of the chosen time slicing scheme.

Next, we consider variability of entropy across all slices, and as before we measure variability by entropy. In order to calculate this higher-order “entropy of entropies”, a suitable binning scheme for entropy values must be chosen, which can be based on the same criteria as outlined above. It is then used to calculate a histogram of the \( ns \) entropy values. We thus define complexity as the entropy of entropy values, which is calculated following Eq. (3),

\[
\text{Complexity} = H(H(X)) = - \sum_{eb=1}^{n_{eb}} p(H_{eb}) \cdot \log_2(p(H_{eb}))
\]

where \( n_{eb} \) denotes the total number of entropy bins, \( eb \) a particular entropy bin, and \( p(H_{eb}) \) the probability of a time slice entropy \( H_s \) falling into bin \( eb \). Complexity measures how uncertain we are about the uncertainty in a particular time slice, when all we know is the distribution of uncertainties (entropies) across all time slices in the time series.

The entire procedure is repeated for many different choices of \( ns \) (time slicing schemes). For each choice of \( ns \), for equal-width slices the width of a time slice is \( sw = nt/ns \). In principle, \( ns \) can be chosen to take any value in the range \([1, nt]\). For \( ns = 1 \), the entire time series is contained in a single slice of width \( sw = nt \). For \( ns = nt \), each time slice contains only a single time step. However, it is recommended to choose \( ns < nt \) - and with it \( sw > nt \) - from a smaller range: If we require that for a robust estimation of a time slice histogram, each of its \( n_{vb} \) bins should on average be populated by a minimum number of \( m \) values, then the width \( sw \) of a time slice (i.e. the number of values within) must at least be \( n_{vb} \cdot m \) (see Eq. 4). This means that for robust estimates of uncertainty, time slices should be wide. For robust estimates of complexity, however, it is the opposite: The histogram of uncertainty values is populated by a total of \( ns \) values (the entropies of all time slices). If we again require that each of the histogram bins should be populated by at least \( m \) values, then at least \( n_{eb} \cdot m \) time slice entropy values are needed. This means that time slices should be narrow to accommodate many of them in a time series of a given length \( nt \) (see Eq. 2). Overall, for a user’s choice of \( m \), Eq. 4 yields the range of recommended time slice widths \( sw \) as a function of time series length \( nt \) and the number of bins for both uncertainty (\( n_{vb} \)) and complexity (\( n_{eb} \)).

\[
\frac{nt}{n_{eb} \cdot m} \geq sw \geq n_{vb} \cdot m
\]

For example, for a time series with \( nt = 30000 \) time steps, and choices of \( m = 3 \), and \( n_{vb} = n_{eb} = 10 \) (all histograms resolved by ten bins), the range of suitable time slice widths is \([30,1000]\).

Throughout all time slicing schemes, the number of value and entropy bins must be kept constant to assure comparability. Together, the set of all time slicing schemes produces a set of complexity-uncertainty value pairs. Plotting them with uncertainty values on the x-axis and complexity values on the y-axis is what we call the complexity-uncertainty-curve, or short c-u-curve. It summarizes several interesting properties of the time series under consideration, which will be discussed in Sect. 3.
3 Application to synthetic and real-world time series

3.1 Time series description

We discuss the properties of the c-u-curve at the example of six time series as shown in Fig. 1a-f. Time series a-c are synthetic time series: a straight line, random uniform noise, and the famous Lorenz attractor (Lorenz, 1963). We chose them for their simple, exemplary and well-known behaviour. The straight line (Fig 1a) contains no variability whatsoever and should therefore show both very little uncertainty and complexity. The random noise (Fig. 1b) contains very high, but constant variability and should therefore show high uncertainty and low complexity. The Lorenz attractor (Fig. 1c) is a prime example of complex behaviour arising from feedbacks in dynamical systems. We used the code as provided by Moiseev (2022) with standard parameters to produce a time series of the Lorenz attractor. From its three variates, for clarity only the first one is shown and discussed, the results for jointly considering all three variates are similar. All synthetic time series consist of \( n_t = 30000 \) time steps, and both for value binning and entropy binning ten bins were used. With a choice of \( m = 3 \), the range of recommended time slice widths is \( s_w = [30,1000] \) according to Eq. 4. In addition to the recommended range of time slice widths, we also included the two extremes values \( s_w = 1 \) and \( s_w = 30000 \) for demonstration purposes.

Time series d-f are hydro-meteorological observations taken from the CAMELS US data set (Newman et al., 2015). The first (Fig. 1d) is daily precipitation observations for the South Toe River, NC (short: STR) basin, the second (Fig. 1e) is the corresponding time series of daily streamflow observations. The basin size is 113.1 km\(^2\), and precipitation mainly falls as rain (fraction of precipitation as snow is 8.5%). The third time series (Fig. 1f) also contains daily streamflow observations, but from the 111.5 km\(^2\) Green River, MA (short: GR) basin, which is more snow-dominated (fraction of precipitation as snow is 22.2 %). We chose the time series for the following reasons: Comparing precipitation and streamflow from the same basin (STR) allows analysing the effect of the rainfall-runoff transformation process on uncertainty and complexity. Here we expect that a basin - by spatio-temporal convolution of precipitation - mainly reduces precipitation variability, and with it uncertainty and complexity. Comparing streamflow from two basins with different levels of snow influence (STR and GR) allows analysing the effect of snow processes on uncertainty and complexity. Here we expect that the carryover effect of snow accumulation, and the influence of an independent additional driver of hydrological dynamics – radiation – should increase both uncertainty and complexity. All hydro-meteorological time series contain 12418 daily observations from 1 October 1980 – 30 September 2014 (34 years). As for the synthetic time series, we also used ten bins to resolve both the range of values and the range of entropies. However, we used a different time slicing scheme to reflect standard ways of time-aggregation of real-world data. In particular, we used the set of \( s_w = \{1,7,14,21,30,60,91,182,365,730,12418\} \) days, which corresponds to 1 day, 1-3 weeks, 1-6 months, 1-2 years, and the entire 34-year period. Please note that for a choice of \( m = 3 \), the range of recommended time slice widths is \( s_w = [30,414] \) days according to Eq. 4. Results for time slices outside of this range should therefore be treated with caution. We included them nevertheless for a more complete assessment of the time series.

We normalized all six time series to a \([0,1]\) value range and then calculated uncertainty and complexity according to Eqs. (2) and (3). Normalizing the data is not a requirement, and it does not affect the shapes and values of the resulting c-u-curves. Rather it was a matter of convenience, as the same binning scheme could then be used for all time series.
Figure 1. Synthetic and hydro-meteorological time series used for demonstration of the c-u-curve. Time series for subplots (a-c) comprise 30000 time steps; for clarity only 300 (subplots (a-b)) and 3000 (subplot (c)) time steps are shown. Time series for subplots (d-f) comprise 12418 daily time steps (34 years); for clarity only four years (1 October 1993 – 30 September 1997) are shown. All values are normalized to [0,1] value range. Further details on the time series are provided in the text.

3.2 Results and discussion

The c-u-curves of all six time series are shown in Fig. 2, and their key characteristics are summarized in Table 1. For clarity, Fig. 3 additionally shows only the hydro-meteorological time series in a sub region of Fig. 2. We start by discussing general properties of the c-u-curve, and then discuss properties of c-u-curves for each time series.

General properties. In Fig. 2 and Fig. 3, both axes are in bit, i.e. independent of the units of the data, thus facilitating intercomparison of different systems, and application to multivariate systems where variates are in different units. Further, for both uncertainty and complexity lower and upper bounds exist. Zero is the lower bound for both, indicating zero within-slice and across-slices variability. Upper bounds arise from the fact that for a given number of bins $n$, the maximum possible entropy is that of a uniform distribution, such that $H_{\text{max}} = H_{\text{uniform}} = \log_2(n)$. For the ten bins chosen here, the upper limits are $\log_2(10) = 3.3$ bit, shown as vertical and horizontal lines in Fig. 2, and horizontal line in Fig. 3. The values of the bounds, and all uncertainty and complexity values of the curve depend on the chosen binning. For direct comparison of c-u-curves, the binning should therefore agree. If this is for some reason not feasible, comparability can be established by normalizing values to a [0,1] range. This is possible due to the above-mentioned existence of upper and lower bounds and achieved by dividing values with their respective upper bound (not shown here). In addition to the existence of general upper and lower bounds, the values of uncertainty and complexity for extreme values of the time-slicing scheme are another general feature of the c-u-curve: For $n_s = 1$, i.e. when the entire time series is contained in a single slice of width $sw = nt$, within-slice entropy is at its maximum and equals the entropy of the time series, and complexity is zero because only a single entropy value populates the entropy distribution. In Fig. 2 and Fig. 3, this limit is indicated by the right end of each c-u-curve. For $n_s = nt$, each time slice is of width one and contains only a single time step. In such a case, within-slice entropy...
is always zero, resulting in zero uncertainty and complexity, i.e. the c-u-curve will always start at the origin (see Fig. 2 and Fig. 3).

Apart from these general features, the overall shape of each c-u-curve contains key characteristics of the underlying time series. We start by discussing the c-u-plot of the straight line in Fig. 2: It shows – as expected - the simplest behaviour: For all time-slicing schemes, both within-slice and across-slices variability is zero, i.e. the series displays zero uncertainty and complexity throughout (all dots are stacked at the origin). As a consequence, mean uncertainty and complexity across all time-slicing schemes (indicated by the brown pentagram in the plot and listed in Table 1) is also zero.

The random noise series in Fig. 2 on the contrary displays very high uncertainty and low complexity for most of the time slicing schemes (most dots are stacked in the lower right corner of the plot), and only for many but narrow time slices of 50, 40 and 30 values per slice does complexity assume non-zero values. This can be attributed to random effects in small samples, where purely by chance both highly and hardly variable samples can occur, thus creating a wide range of time slice entropies, resulting in apparent non-zero complexity. For wider slices, the larger sample size leads to more similarly distributed samples, resulting in a narrow range of time slice entropies and hence low complexity. Overall, mean uncertainty is very high and mean complexity is very low (position of the pink pentagram in Fig. 2 and values in Table 1), which is what we expected from random noise as a purely chaotic process.

The Lorenz attractor in Fig. 2 reveals a more diverse behaviour across the range of time slicing schemes. We start discussing it for the case of $sw = 30000$, i.e. when a single time slice covers the entire time series. As described in the general

Figure 2. C-curves for synthetic (dotted) and hydro-meteorological (no marker) time series as shown in Fig. 1. Time series length is 30000 for the synthetic data and 12418 for the hydro-meteorological data. The number of value bins and entropy bins is ten, maximum uncertainty limit and maximum complexity limit is at $\log_2(10) = 3.32$ bit. For the synthetic series, dot labels indicate the time slice width $sw$ used to calculate uncertainty and complexity, and the pentagram positions indicate mean uncertainty and mean complexity across all chosen time slicing schemes. The hydro-meteorological series are included to indicate their position within the full range of uncertainty and complexity; their details are shown in Fig. 3.
properties, for this case uncertainty is always at its maximum and equals the entropy of the time series, and complexity is zero, because only a single entropy value populates the entropy distribution. The actual uncertainty value (3.11 bit), or its distance from the upper limit of uncertainty (3.11/3.32 = 94%), is a key characteristic of the time series and expresses its overall variability. Decreasing the time slice width $s_w$ decreases within-slice variability (uncertainty). Also, it provides the potential for nonzero complexity as more and more entropy values populate the entropy distribution. For the curve shown in Fig. 2, complexity continuously increases and reaches its first maximum value of 2.66 bit (or 2.66/3.32 = 80%) for $s_w = 200$ and at 2.22 bit of uncertainty. This point is another key characteristic of a c-u-curve, indicating at which temporal aggregation the across-slices variability is highest. Further decreasing slice width first leads to a decrease and then another increase in complexity until a second maximum of 2.66 bit is reached at $s_w = 50$ (see values in Table 1). Afterwards, complexity and uncertainty decrease to zero for $s_w = 1$, which is a general property of any c-u-curve (see discussion of general properties above). Taking the uncertainty and complexity mean across all time slices summarizes the c-u-curve in a single point (purple pentagram in Fig. 2, values in Table 1). For the Lorenz attractor, it reveals medium average uncertainty, and high average complexity. This is in accordance with expectations, as the Lorenz attractor is known for exhibiting complex behaviour. Interestingly, apart from revealing its generally complex behaviour, the c-u-curve also reveals at which particular time slice width complexity of the Lorenz attractor is at a maximum. This can be interpreted as a “characteristic time scale” of the time series.

Next, we discuss the c-u-curves of the hydro-meteorological time series. In Fig. 2, they are indicated by the lines without markers. It is immediately obvious that they all possess low uncertainty, much lower than the theoretical maximum (indicated by the vertical “max. Uncertainty” limit) and the random noise, and also lower than the Lorenz attractor. This is in accordance with our expectations, and a consequence of the typically high temporal autocorrelation of hydro-meteorological time series, which clearly separates them from purely random time series. For a better view of detail, we re-plotted the hydro-meteorological time series in a sub region of the uncertainty limits in Fig. 3, which we will refer to in the following.

Despite the generally low uncertainties, the precipitation STR time series in Fig. 3 displays considerable average complexity (indicated by the pentagram position), which can be explained by the existence of meteorological regimes with different levels of precipitation variability, such as dry periods (low variability), periods with alternating dry and wet periods (high variability), and wet times with diverse precipitation amounts (high variability). The highest complexity occurs for a time slice width of $s_w = 14$ days, indicating that the greatest variability of within-slice precipitation variability occurs for two-week periods.

Interestingly, the corresponding streamflow STR time series displays much lower mean and maximum values (see Table 1) for both uncertainty (within-slice variability) and complexity (across-slices variability). This is in accordance with the general hydrological understanding that in the absence of major carryover mechanisms, rainfall-runoff transformation in

<table>
<thead>
<tr>
<th>Time series</th>
<th>Uncertainty (bit)</th>
<th>Complexity (bit)</th>
<th>Characteristic time scale (days)*</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>max</td>
<td>mean</td>
<td>max</td>
</tr>
<tr>
<td>Line</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Random noise</td>
<td>3.32</td>
<td>3.04</td>
<td>0.67</td>
</tr>
<tr>
<td>Lorenz attractor</td>
<td>3.11</td>
<td>1.84</td>
<td>2.66</td>
</tr>
<tr>
<td>Precipitation STR</td>
<td>0.38</td>
<td>0.30</td>
<td>1.49</td>
</tr>
<tr>
<td>Streamflow STR</td>
<td>0.09</td>
<td>0.06</td>
<td>0.53</td>
</tr>
<tr>
<td>Streamflow GR</td>
<td>0.80</td>
<td>0.57</td>
<td>2.33</td>
</tr>
</tbody>
</table>

* width of time slice at which maximum complexity occurred.

Table 1. Key characteristics of the c-u-curves for both the synthetic and the hydro-meteorological time series.
catchments is mainly by aggregation and convolution, thus reducing the variability of the precipitation signal. It is noteworthy that while this harmonizing effect changes uncertainty and complexity means and maxima, it does not affect the characteristic time scale: For streamflow STR - just as for precipitation STR - it is two weeks. This suggests that precipitation remains the main control of streamflow complexity, despite the processes involved in rainfall-runoff-transformation.

This is different for the second streamflow GR time series. Here, in addition to the above-mentioned rainfall-runoff transformation, precipitation is partly stored as snow and later released as streamflow by melting. The temporal pattern of snowmelt is not only governed by snow availability, i.e. the precipitation regime, but also energy availability, i.e. the long-term radiation and temperature regime. Such additional, independent controls of hydrological function can add uncertainty and complexity to streamflow production. Compared to streamflow STR, both uncertainty and complexity are indeed much larger in terms of mean and maximum values, they are even larger than the corresponding values for precipitation STR (compare pentagram positions in Fig. 3 and values in Table 1). The characteristic time scale of streamflow GR is at 2-3 months (60-91 days). This is considerably longer than for streamflow STR, and can be explained by the carryover effect of snow accumulation and snowmelt acting at time scales in the order of months rather than days or weeks.

![Figure 3](https://example.com/figure3.png)

Figure 3. C-u-curves for all hydro-meteorological time series as shown in Fig. 1d-f. All time series comprise 12418 time steps, the number of value bins and entropy bins is ten, maximum uncertainty limit and maximum complexity limit is at \( \log_2(10) = 3.32 \) bit. Note that for better display of details this is a horizontally zoomed-in version of Fig. 2. Dot labels indicate the time slice width \( s_w \) used to calculate uncertainty and complexity. The pentagram positions indicate mean uncertainty and mean complexity across all chosen time slicing schemes.

4 Summary and conclusions

In this paper we presented a method to analyse and classify dynamical systems by the two key features uncertainty and complexity. After dividing the time series into a set of time slices, the Shannon information entropy is calculated for the data
in each time slice. Uncertainty is then calculated as the mean entropy of all time slices, complexity as the entropy of all entropy values. Complexity thus expresses “uncertainty about uncertainty” in the time series. Calculating and plotting uncertainty and complexity for many time slicing schemes yields the c-u-curve, with key characteristics mean and maximum uncertainty, mean and maximum complexity, and the characteristic time scale of the time series. The latter is defined as the time slice width at which maximum complexity occurs.

The c-u-curve method has several useful properties: Independence from the units of the data (both uncertainty and complexity are expressed in bit), existence of upper and lower bounds for both uncertainty and complexity as a function of the chosen data resolution, and bounded behaviour when approaching upper and lower limits of time-slicing: For a single time-slice containing all data, uncertainty equals the time series entropy and complexity is zero, for time-slices containing single values both uncertainty and complexity are zero. The c-u-curve method is applicable to single- and multivariate data sets, and to deterministic and probabilistic value representations (ensemble data sets), making it suitable for a wide range of tasks and systems. The main limitation of the method arises from the requirement of sufficiently populating distributions, which sets bounds to both the minimum and maximum width of time slices.

We provided a proof-of-concept at the example of six time series, three of them artificial, three of them from hydro-meteorological observations. The artificial time series (straight line, random noise, Lorenz attractor) were chosen for their very different, exemplary and well-known behaviour, and with the goal to demonstrate that the c-u-curve successfully reveals this behaviour, i.e. to demonstrate the general applicability of the method across a wide range of time series types.

The observed time series (precipitation and streamflow from a mainly rainfall-dominated basin, and streamflow from a basin where additionally snow processes influence the hydrological function) were chosen with the goal to demonstrate that the c-u-curve method reveals characteristics of real-world time series that are in accordance with general knowledge of hydrological system functioning. For all time series, we could show that the c-u-curve properties were distinctly different among the time series – which indicates that the method has discriminative capabilities useful for system classification -, and that the properties are in accordance with expectations based on system understanding – which indicates that the method captures relevant time series properties and expresses them in terms of uncertainty and complexity -. While the range of applications presented in this paper is small, and mainly intended as a proof-of-concept, the results encourage further studies. Particularly for hydro-meteorological applications, we suggest that the c-u-curve method can be used for data-based system classification, which is a precondition for the important hydrological task of regionalisation, i.e. the transfer of knowledge from well-observed to poorly-observed places. In this context, the c-u-curve and its characteristic values can be used in a fashion similar to current uses of the flow-duration curves. This is supported by the clear differences of c-u-curve properties between the two investigated streamflow time series. We further suggest that the c-u-curve and its characteristic values can be used as a useful additional objective function in hydrological model training: While standard hydrological objective functions such as Nash-Sutcliffe efficiency guide models towards point-by-point agreement of models output and observations, c-u-curve characteristics can guide models towards correct representations of short- and long-term variability patterns. Further work on this topic is in progress.

**Code availability.** The code used to conduct all analyses in this paper is publicly available at [https://doi.org/10.5281/zenodo.5840045](https://doi.org/10.5281/zenodo.5840045) (Ehret, 2022).

**Data availability.** All data used to conduct the analyses in this paper and the result files are publicly available at [https://doi.org/10.5281/zenodo.5840045](https://doi.org/10.5281/zenodo.5840045) (Ehret, 2022).

**Author contributions.** UE developed the c-u-curve method and wrote all related code. UE and PD designed the study together and wrote the manuscript together.
Competing interests. The authors declare that they have no conflict of interest.

Acknowledgements. We gratefully acknowledge support by the Deutsche Forschungsgemeinschaft (DFG) and the Open Access Publishing Fund of the Karlsruhe Institute of Technology (KIT). The second author acknowledges the funding received from the research project “Advanced Hydrologic Research and Knowledge Dissemination” sponsored by the Ministry of Earth Sciences (MoES), Government of India (Project No. MOES/PAMC/H&PC/41/2013-PC-II).

Financial support. The article processing charges for this open access publication were covered by a Research Centre of the Helmholtz Association.

References


