Hydrology and Earth System Sciences Discussions

Hydrol. Earth Syst. Sci. Discuss., referee comment RC1
https://doi.org/10.5194/hess-2022-16-RC1, 2022

Comment on hess-2022-16

Jasper Vrugt (Referee)

Referee comment on "Technical note: Complexity–uncertainty curve (c-u-curve) – a method to analyse, classify and compare dynamical systems" by Uwe Ehret and Pankaj Dey, Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2022-16-RC1, 2022

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

Summary: The authors resort to information theory and present the so-called c-u-curve to describe/quantify characteristic properties of hydrometeorological data. This c-u-curve displays graphically the relationship between what authors refer to as system uncertainty and complexity. This information is thought to be expressed and/or contained within spatial and/or temporal measurements of the data generating process of interest. **System uncertainty** is defined as the mean Shannon information entropy of many different time slices (time windows). The authors define **system complexity** as the '*…uncertainty about uncertainty*' (P1, Line 11) and express this quantitatively as the entropy of the entropies of all time slices. As the two metrics depend strongly on the temporal extent (width) of the time window, the authors repeat their analysis for many different slice sizes. The c-u-curve is a graphical depiction of the relationship between the so-obtained system uncertainty (*x*-axis) and system complexity (*y*-axis), both of which have units of bits. The authors illustrate this idea by application to six different signals (time series), including simulated data of a (i) deterministic (horizontal line), (ii) random (normally distributed variates) and (iii) chaotic system (Lorenz attractor) and measured time series of (iv) precipitation and (v,vi) catchment discharge of the South Toe and Green Rivers in the United States. The authors conclude that the c-u-curve can be used to analyze, classify and compare dynamical systems.

Evaluation: The manuscript discusses an important topic in hydrology and complex systems analysis in general, namely the characterization of the dimensionality and complexity of dynamical systems. I enjoyed reading this manuscript. The document is well written and relatively easy to understand. Rationales and ideas are clearly presented. The six case studies demonstrate/showcase the potential use of the c-u-curve, inform readers about the methodology and how to interpret its results. I applaud the authors for their work, which I believe is very interesting. I do have serious concerns however about the mathematical and literature underpinning of the methodology, and the robustness and convergence properties of the c-u-curve. Based on these comments, I recommend a **major revision**.

General comment

The authors decided to present their work in the form of a technical note. This is an efficient way to rapidly disseminate new ideas. But technical notes have strict length requirements which can make it difficult to address all important aspects of the work presented. The ideas presented are very interesting, yet a full paper may do more justification to the ideas and work presented. I have several questions about the methodology, which I think should be addressed before readers can judge that what is presented is a substantial and/or important advance in our ability to analyze, classify and compare dynamical systems. Note that in my review below I use the word 'signal' for a measured or simulated time series of some quantity of interest. I also use the word 'paper' in reference to this technical note. This word is conveniently used and should certainly not imply that I was expecting a much longer manuscript.

Specific comments

- Section 2.1: The authors resort to information theory to analyze the temporal properties of the signal of interest. They coin two measures of system functioning/behavior, namely system uncertainty and system complexity and provide a mathematical definition for both that are subsequently used to construct the c-u-curves in the analysis. No references and/or background is provided about their definition. There is a large body of literature on complex systems and imagine that others have defined similar metrics to classify, describe and characterize time series data. This begs the questions whether the two criteria stand completely on their own and if earlier attempts have been made to analyze time series data in a similar fashion? I think the paper would be considerably stronger if the authors can relate their work to previous published work. Have other definitions of these criteria appeared in the complex systems literature?
- Section 2.1: Related to my previous comment, what is wrong with specifying system complexity as the temporal variance of the Shannon entropies? Then you the first moment (mean) as measure of system uncertainty and the second moment (variance) as measure of system complexity. I am sure the authors have thought about this. In the present paper I just miss rationales and arguments so as to why their definitions are appropriate – also in light of past work done in the literature on this topic.
- Section 2.1: In their analysis of the signal, the authors coined the words system complexity and system uncertainty. I do not think these labels are accurate. System uncertainty and system complexity refer to the system as a whole – and should, in principle, not depend on which variable of the system is observed. They should be invariant properties (unless the system experiences change). Instead, what the authors determine is the uncertainty and complexity of the signal only. Thus, I think it is more accurate to use the words signal uncertainty and signal complexity. Indeed, I expect you will get different c-u-curves for different signals of system behavior. If we take hydrology as example, then soil-moisture will likely give a different c-u-curve than a time series of groundwater table depths and this curve will be different from its counterpart of the discharge. Certainly, I would argue that a single component of system behavior is insufficient to characterize the complexity and uncertainty of the system as whole.

- Section 2.1: The choice of the number of time slices and their spatial extent; I'll call this the temporal discretization of the signal; play a crucial role in the analysis. Without derivation and much explanation at all the authors introduce Equation (4) which provides a lower and upper bound for the width of the time slices. How is this equation derived? Is this a rule of thumb? The lack of a theoretical underpinning is a concern. It may be productive to have a look at Sturges method (or for that matter Scott's method or Freedman-Diaconis) which provide a rule of thumb for the number of histogram bins that should be used for a given length of data. This may be used to improve the statistical underpinning of Equation (4).
- Section 2.1: Readers may be interested to see a few of the histograms that went into the computation of system uncertainty and system complexity. Do the histograms differ substantially from one time slice to the next? Do they have an overarching distribution? Skew, kurtosis, etc?
- Figures 2 and 3: The authors assume that the c-u-curve is continuous, and connect the individual (u,c) pairs for individual time slices with a solid line. But is the curve continuous? Are there theoretical arguments from which one expects the curve to be continuous and not discrete?
- Figure 2: What happens to the Lorenz c-u-curve if we use windows (slices) of a size smaller than 30? Equation (4) suggests that such value is not recommended, but what happens to the curve itself? Does the curve oscillate close to the origin?
- Liner 145: The authors use normalization of the signal to yield values between 0 and 1. This itself is inconsequential yet allows a fixed recipe for data types with very different magnitudes. Why do the authors not use a similar normalization in the time domain? This may help (or not) to standardize the characterization of the width of the slices. The only variable left is then the number of data points.
- How does the c-u-curve respond to the frequency of measurement of the signal? For example, in the case of discharge, you can construct the curve for hourly, daily, weekly and monthly data (average flows) – do we see convergence of the c-u-curve to it counterpart of the horizontal line? I expect such convergence to be faster for time-average data points than for a signal made up of instantaneous measurements. This analysis is important as it will help establish convergence properties of the c-u-curve.
- How does the c-u-curve respond to a) numerical errors (in case of simulated signals) and b) system nonstationarity? This analysis is not difficult to do with simulated discharge data (for example using fixed step integration versus a variable time step or implicit solution) and will provide further insights into the method.
- What is the effect of data transformation on the inferred c-u-curve? The authors can again resort to discharge data and compare the c-u-curve of the original signal with its counterpart derived from a Box-Cox transformation. One could even consider wavelet analysis, but this is for future work.
- Equation (4): I find the choice of mathematical variables not particularly intuitive. The authors must have thought about their choice of symbols much better than I did, but why not assume at the outset that we are looking (typically) at temporal data and use Dt for the temporal extent of the time slice, n, for the number of slices and so forth. Then one can assign subscripts to these variables to differentiate between their definitions for the two metrics.
- Line 97: Why should each bin of the histogram have at least some nominal number of m values? This seems rather artificial. Why not construct the histogram using the rules of thumb of Scott or Sturges? A bin cannot have zero values as this introduces difficulties with the computation of the density and log of the density in Eqs. (1) and (3)? I can only recommend googling '*How to calculate the Kullback-Leibler divergence for discrete distributions*' – this will provide ways forward how to compute the product of p_i and log(p_i) if p_i is zero. Additionally, the authors can think of a Gaussian mixture model to fit a distribution through the histogram and use this fitted mixture to compute system uncertainty and system complexity. This process is sufficiently fast to warrant practical use in a long time series with many slices.
- Line 196: The authors refer to random noise as a purely chaotic process. I do not think

this is an accurate description of random noise as draws from a normal variate do not satisfy the definition of chaos.

- The authors analyze the daily discharge data of two watersheds. First, I believe that hourly data is available for most watersheds. Maybe not for > 20 years uninterrupted but for sufficiently long times to satisfy the requirements of the methodology. More importantly, as a demonstration of the power and usefulness of the methodology the authors should consider using watersheds of the CAMELS data set with different hydrologic regimes (see recent classification methods published in HESS). How does the c-u-curve respond to the hydrologic regime? Do we see differences among all hydrologic regimes? And if we take multiple watersheds from the same regime, do they group in the c-u space? I consider this extension to all (4 or 5) hydrologic regimes to be important as it tests the robustness of the methodology.

I very much enjoyed reading this paper. From my comments above it is clear that I have concerns about the statistical/mathematical and literature underpinning of the methodology, the use of the words system uncertainty and system complexity, and the robustness and convergence properties of the methodology (= c-u curve). The additional studies I suggest will help answer important questions about the usefulness and diagnostic power of the c-u curve and its use in the analysis and classification of hydrometeorological time series. My comments are intended to help the authors further refine/improve their methodology for maximum exposure and use in the community.

PS. I did not proofread my review. Also, my comments are listed in a somewhat random order (with a c-u-curve that approaches a point) as a result of going back and forth in the paper.

Jasper Vrugt