hess-2022-16 Responses to comments by the editor (Jim Freer) as of 28 Feb 2023

Editor comments are in blue, our replies are in black.

Public justification (visible to the public if the article is accepted and published):

Dear Authors,

Whilst the reviewers appreciate the changes within the manuscript and that there is merit in the research presented they still note some difficulties (in their view) with the readability, a slightly dense and convoluted manuscript at times especially when introducing the core methods, and most importantly a continuing discussion about ensuring the theoretical underpinnings are clear and dealt with appropriately in the manuscript. I accept that for a technical note paper we need to ensure a strong connection to the theory and previous research that is related and I agree with the points that particularly the 1st reviewer makes. These reviewers are both considerable experts in their field. So I request again the authors consider these points raised from their revision but I do think the discussion is still at a phase where we will need to ask at least one reviewer for confirmation of how the manuscript has been changed to reflect these continuing points on the theory surrounding this research. I look forward to the authors changes to the manuscript and detailed responses to the points raised before completing one final review round, best wishes, Jim

Additional private note (visible to authors and reviewers only):

Dear Authors. I think you are improving the manuscript but we are not there quite yet. So please ensure that you respond to all the comments presented in the revision round and how you can ensure the clarity in your methods and how this relates to prior theory and developments. We do need to get that right for a technical note. I completely appreciate this can be a challenge when we have such a broad set of related research but I would still say there are some core discussion points that are being brought up that need more attention. The point here is now to either defend your position or think of how to better weave into your developments the research and nomenclature that has come before and/or is related - and thus how your methods sit and are different and why therefore they have new insights. I will be sympathetic to any strong justifications as to why you can defend your approach to your presentation but these still need to be answered. I am in agreement with your defence of this being a technical note, so I don't expect a change in the submission type, best wishes, Jim

Dear Editor,

We have updated our manuscript to improve readability according to the referees suggestions. Please see our related replies in the point-by-point replies to the referees. In terms of the questions about theoretical unterpinnings of the c-u-curve method by referee #,1, Jasper Vrugt: We have recently had the opportunity to listen a presentation of his recent work, and had the opportunity for an in-depth discussion afterwards. From that it became clear the referee focuses on questions of comparing one distribution (e.g. from a candidate model during calibration, or from a forecast) against a reference, typically observed evidence). In this context, the referee suggestions about the use, and decomposition of KL-divergence make sense. However, the goal of the c-u-curve method is on characterization of a single data set in terms of internal uncertainty and complexity, where entropy rather than KL-divergence applies. We also replied to the referee question about the connection of the law of total expectation and the c-u-curve method (comment 2): The short answer here is that we do make use of the total law of expectation, but for the (common) special case of uniform slice width the calculation of uncertainty simplifies from the equation shown in our reply to comment 2, to the E1. 2 in the manuscript. We would like to point out that overall, both of the

referees requested only minor revisions, and we hope that our related changes to the manuscript, and our point-by-point justifications where we decided to not change it, will be sufficient to make it acceptable in the current form.

Yours sincerely,

Uwe Ehret and Pankaj Dey

hess-2022-16 Responses to comments by referee #1 (Jasper Vrugt) (report #1 as of 12 Jan 2023)

Dear Editor, dear Referee,

We have revised our manuscript based on the comments by the first referee, Jasper Vrugt, along the lines of our replies to the referee. In the following we will repeat the comments (in blue) together with the replies (in black). We also indicate for each comment the lines in the manuscript (in red) where we applied the changes. The line numbers are for the revised version in track change mode.

Comment 1: I enjoyed reading the revised manuscript. As I said in my original review, the work presented is exciting and brings up lots of new ideas that can be explored in future work. The revision satisfactorily addresses most of my original review comments, yet, I am not certain that a technical note does complete justice to the material presented. The material is interesting and important but the revision has led to a dense manuscript; I thought the original submission was easier to read. Also the construction/definition of the c-u-curve bring up other important technical (implementation) and theoretical questions that are not fully explored. This can be resolved through revision or done in future research using the present paper as reference material. I am leaning towards the second option - in part also because software is provided in a separate Zenodo DOI. The only down-side is that future research may make certain aspects of the approach incomplete/obsolete or subject to redefinition, in response to a proper theoretical foundation and redefinition of uncertainty and complexity. See my comments below.

Reply 1: We are glad about the overall positive evaluation of our revisions by the referee. We agree with the referee that this manuscript is at the edge between a technical note and a scientific paper. We have decided to publish the core method as a technical note for the reasons explained in the first round of reviews, and to present applications in a follow-up research paper. We agree with the referees preference to publish further developments of c-u-curve theory in future papers, with reference to this manuscript, because we think this manuscript presents a coherent piece of research, and to ensure timely publication. This decision comes with the need for brevity for the technical note. Mostly based on the suggestions of the referees in round one, we have revised the manuscript, which led to a substantial increase in length (e.g. the added literature overview in the introduction, the discussion of entropy vs. variance, the additional appendix with exemplary histograms, etc.). We are sure these additions have strengthened the manuscript, but at the same time we have to keep an eye on manuscript length. In terms of a potential need for a later redefinition of c-u-curve aspects, please also see our replies to comment 2.

Theoretical questions that deserve attention/further research:

Comment 2: 1. Now that I have accidentally educated myself a bit further on the topic of information theory as part of an article that I was working on (different topic) I would recommend the authors to look into a deeper theoretical foundation for their c-u-curve. Is it not possible to find mathematical/statistical expressions for the mean of the entropy - and the entropy of the entropy using the law of total expectation (with the entropy, H(X), and data, X, that are random variables on the same probability space). See what this decomposition brings. This is the theoretical underpinning in continuous space and leads to a discretized form.

Reply 2: The law of total expectation is used in the c-u-curve method when calculating uncertainty as the expected value of entropy (Eq. 2). In other words, what is done in Eq. 2 is that when we are asked to give a single-valued best guess of the entropy within any time slice, then the expected value of all time-slice entropies is exactly this best guess. If the time slice widths differ, then we need to consider the occurrence probability of each slice when calculating the expected value, and the width of each time slice (nt) relative to the total length of the time series (T) would be an appropriate measure of

this occurrence probability. In this general case, the law of total expectation would be used to calculate uncertainty:

$$E\big(H(X)\big) = \sum_{s=1}^{ns} H_s(X|s) \cdot p(s), where\ p(s) = \frac{nt(s)}{T}$$

In the case that nt of all slices are equal (equal-width slices), this equation simplifies to Eq. 2 as shown in the manuscript. This is also mentioned in the manuscript in line 114. In this sense we would like to respond to the referee that his suggestion is already incorporated in the c-u-curve method.

Comment 3: 2. The authors define the mean (expectation) of the entropy as uncertainty. This definition is at odds with the common definition of uncertainty in forecasting systems - which defines uncertainty as the entropy of the mean (probabilities).

Reply 3: We termed expected entropy as "uncertainty" as it specifies, in a single number, how uncertain we are on average (over all time slices) when we have to guess a particular value within a particular time slice. Due to its formulation in information terms, this uncertainty has a very intuitive interpretation of "number of binary questions to ask, if the distribution (of values within the time slice) is known and one after another all values in the time slice have to be guessed". Multiplying the uncertainty as we calculate it in Eq. 2 with the total number of points nt of the time series yields exactly the total number of questions needed to be asked to guess all values in the time series, if for each value we know the distribution of the time slice it is in. Interpreting entropy as a measure of uncertainty is at the very heart of information theory and was not invented or coined by us, and it is to our knowledge not add odds with the use of uncertainty in forecasting systems, where the uncertainty of a particular forecast can be measured by the spread of the prediction ensemble via the entropy of the ensemble.

Comment 4: 3. I recommend the authors to have a look at the decomposition in Weijs et al. (2010) of the KL-divergence. This is based on earlier work of Murphy (1973) on decomposition of the Brier Score. This results in 3 terms; uncertainty, reliability and resolution.

Reply 4: After the referees presentation of his recent work in the "paper club infotheory", we now understand much better what he means by this comment. One focus of the referees work is on evaluating probabilistic forecasts against evidence (typically crisp, but could also be probabilistic), and for this purpose KL-divergence is an appropriate measure, and it can be decomposed as indicated by the referee. However the purpose of the cu-curve method is not the comparison of one data set versus another, but characterization of a single data set in terms of internal uncertainty and complexity. We therefore agree with the referee that it is worth looking at the decomposition of KL-divergence, but that this does not specifically apply to the cu-curve method.

Comment 5: 4. The theorem/proof the authors provide on P.14 is mathematically/statistically delinquent. First, the authors talk about a probability distribution "p" on S. This should be a probability measure (sums to 1). This measure should be part of a convex class P and be quasi-integrable with respect to all P? Formula B2 warrants a reference.

Reply 5: As indicated at the beginning of Appendix B, we here as closely as possible repeat Theorem 5.12 from Conrad (2022), and therefore also adopted the notation of Conrad (2022), where p is a discrete probability distribution on p_j possible states. We suggest keeping to this notation for easier linkage to the rest of Conrad (2022). Also, as we make clear that all of Appendix B2 just repeats Conrad (2022), we suggest that it is clear for the reader that formula B2 is also from that source.

Comment 6: 5. In the comparison to existing methods the authors bring up the KL-divergence. This is equal to the divergence of the logarithmic score, which in turn has as its generalized entropy function

negative Shannon entropy. This is the theoretical link between what is presented in the paper and Lopez-Ruiz.

Reply 6: Using KL-divergence as a measure of disequilibrium between the system and a maximum entropy benchmark was introduced Feldman and Crutchfield (1998), replacing the sum of squared differences used by Lopez-Ruiz (1995). Also, we would like to point out that there are fundamental differences between the complexity measures proposed by Lopez-Ruiz (1995), Feldman and Crutchfield (1998) and our approach, see lines 250-255 in the manuscript: "… but the essential differences of CLMC and the c-u-curve methods remain: Firstly, the former defines complexity as the product of two separate system characteristics, of which one is the departure from a benchmark system, the latter derives both characteristics from the system alone. Secondly, the former does not take the order of the data into account, while the latter explicitly does when calculating entropy for data within temporally neighbouring data within time slices."

Comment 7: 6. If the authors generate ensembles with the Lorenz attractor by assuming some distribution for the parameters. This would yield a distribution forecast from which the entropy can be computed. This provides an entropy at each time; how does this relate to the averaged entropy of the time slices? And how would the c-u-curve look?

Reply 7: We agree that it would be interesting to apply the cu-curve method to probabilistic time series, e.g. coming from the Lorenz attractor applied with a distribution of parameters, or from an ensemble weather forecasting system. This is indeed possible, and we provide examples thereof in the example applications in Ehret (2022), but for brevity do not present such an example in the paper (we mention this now in more detail in lines 223-229). We will do so in the follow-up research paper we are currently working on. Indeed we think it is one of the strengths of the method that the extension to multivariate and probabilistic cases is seamless: When moving from univariate to multivariate cases, the entropy within a time slice simply changes from uni- to multivariate entropy. When moving from deterministic to probabilistic variables, for each time step in a time slice, a value distribution rather than a crisp value will be used to populate the distribution of all values in the time slice, but the result will still be a single distribution with a single entropy value, which can be plotted as before in the c-u-curve. If the spread of the ensembles at each point in time is large compared to the spread of values over time within the time-slice, the overall time-slice entropy will be dominated by the ensemble spreads. In short, uncertainty of the c-u-curve method measures uncertainty in time as well as uncertainty due to ensemble spread at one point in time.

Technical/presentation questions

Comment 8: 1. I do not find the use of symbols intuitive. Each time when I read "nbv" I think this involves multiplication of n, b and v. Also, strictly speaking if variables are acronyms themselves they should be written in regular script? I leave this to the authors, but personally would prefer just picking regular symbols.

Reply 8: We agree with the referee that "nvb" can be mistakenly interpreted as a product of three variables, and that this misinterpretation should be avoided. We do so in the manuscript by using the "·" symbol each time a multiplication is done. For clarification, we have added a related explanation at the beginning of section 2.1 (lines 90-93). We also considered using subscripts for all but the first symbol (e.g. "n_{vb}" instead of "nvb") to avoid confusion with multiplication, but this partly leads to double-subscripts which are hard to interpret as well (e.g. x_{vb} in Eq. 1 would become x_{v_b}). Also, we consulted fellow mathematicians on the appropriateness of our use of symbols for variables. Their feedback was that the use of symbol combinations for variables (such as "nvb") is not standard, but fully acceptable as it is an unambiguous notation, and because the use of the symbols is explained at the beginning of section 2.1, and each variable is explained in the text directly where it was introduced in an equation (e.g. "nvb" is explained below Eq. 1). We therefore prefer to keep the use of symbols as is.

Comment 9: 2. A technical note has a limited number of words. As a result, the authors have to be short in their description of the case studies. Personally, I think the paper would be easier to read if those details were presented. For instance, the Lorenz attractor. The authors refer to a code (Line 256), but further details are missing. As a result, on Line 257-258 the write that "From its three variates … only the first one is shown…". Not all readers may be familiar with the Lorenz 1963 model - nor with the computational implementation, which, by itself is not trivial (different codes do not always produce the same chaotic behavior!!); certainly they may not know the three variates, etc. Of course the software will help. But it illustrates my struggle of wanting to accept this paper with minor revision, but at the same time wanting to make sure there is enough detail for readers to understand what is presented.

Reply 9: For the general topic of "technical note" or "scientific paper", and the need for brevity, please see our response to comment 1. About the Lorenz attractor in particular: While a more in-depth explanation of the Lorenz attractor would be interesting, we argue that it is not indispensable for this manuscript, because we only use the time series as a nice example of a complex time series. The complexity of the series is directly visible from Fig. 1(c), and no deeper understanding of the Lorenz attractor is required for this. For the interested reader, we provide both the reference to the original paper (Lorenz, 1963) and the code and settings we used to generate the data (Moiseev, 2022), which we think balances the need to inform the reader with the need for brevity.

Comment 10: 3. Is unit "bit" or "bits"?

Reply 10: Thanks for raising this point, we realized we were inconsistent with the use of "bit" and "bits". The unit is "bit", and whenever a reference is made to the unit in general, the singular is used (e.g. "entropy is measured in bit"). Whenever a reference is made to the entropy of a particular quantity, the plural is used (e.g. discharge has an entropy of 5.3 bits"). We have checked the manuscript once more for consistent use of "bit" or "bits", and updated it where necessary.

Comment 11: 4. Line 462: replace computer with iterative algorithmic recipe?

Reply 11: Thanks. We modified the sentence to "The value of β can be numerically approximated with an iterative algorithmic recipe and Eqs. B1 and B2 (see example 5.14 in Conrad, 2022)." See lines 478-479.

Comment 12: 5. Line 455: with given \overline{E} and maximum entropy?

Reply 12: Agreed. We removed "having" from the sentence (line 472).

Comment 13: 6. Equation (1): "X" -> is not defined; the sample data within the bin/slice. Or "X" is the random variable of interest of which small x are the samples? Also should log_2 not be upright as it is a mathematical function?

Reply 13: Thanks. X is indeed the entire sample data in the slice, and x are the subset of X falling into a particular value bin. We have added a definition of X in line 101.

Log_2: Thanks, we changed it to upright, also in Eq. (3)

Comment 14: 7. Line 98: If slices are not of uniform width - would this not create difficulties with entropy comparison/averaging, etc.?

Reply 14: Yes, in such a case calculating uncertainty according to Eq. 2 should consider the relative contribution of each slice (by weighting its entropy with the slice width relative to the total length of the time series). Surely in most cases uniform slice widths are preferable, the point we wanted to make with our statement is that choosing non-uniform slice widths is not impossible per se.

Comment 15: In summary, I enjoyed reading this revision - I have stated my concerns (that is why I rate the manuscript as good instead of excellent), but believe that the ideas presented are worthy of a prompt publication.

Reply 15: Again we thank the referee for the overall positive evaluation of our manuscript, and appreciate his comments and suggestions that led to a substantial improvement of the manuscript.

hess-2022-16 Responses to comments by referee #2 (report #2 as of 18 Jan 2023)

Dear Editor, dear Referee,

We have revised our manuscript based on the comments by the second referee along the lines of our replies to the referee. In the following we will repeat the comments (in blue) together with the replies (in black). We also indicate for each comment the lines in the manuscript (in red) where we applied the changes. The line numbers are for the revised version in track change mode.

The revised paper has addressed the major concerns raised in previous round of reviews. I have a few relatively minor remaining concerns:

Comment 1: 1. In the introduction, around lines 35-40 it is stated that "Interestingly, despite its importance and widespread use there is to date no single agreed-upon definition and interpretation of complexity". The newly added references there are helpful, but what is missing is at least some brief summary of what these references actually say. Such summary would help support the earlier statement.

Reply 1: Agreed. We have added brief summaries of Gell-Mann (1995), Lloyd (2001), Prokopenko et al. (2009) and Ladyman et al. (2013) to the introduction in lines 39-46.

Comment 2: 2. The presentation around lines 85-90 is rather convoluted and confusing. For example:

- it starts by citing the Xenodo repository (which is essentially supposed to serve as a repository and appendix to the paper) where the method is generalized before actually being presented in the paper itself.

- there are various sentences such as "Also, we calculate discrete entropy based on a uniform binning approach" the significance of which is unclear based on the presentation thus far.

- "Nevertheless, the method can also be used with non-uniform binning or continuous representations of data-distributions" - I am not sure how acceptable is this claim without actual proof/demonstration, especially once again before the method itself is even presented.

I would recommend that early part of the presentation be revised for clarity, e.g. that statements / generalizations that only make sense after the presentation is complete be moved to the Discussion section, speculative statements be removed or identified as such, etc.

Reply 2: Agreed. We moved most of this introductory section to the discussion of properties in section 2.2 (lines 223-238), such that the method is first introduced, and then the generalizations and limitations are discussed. We also added some more explanations to the text.