Hydrol. Earth Syst. Sci. Discuss., 10, C1665–C1671, 2013
www.hydrol-earth-syst-sci-discuss.net/10/C1665/2013/

Hydrology and Earth System Sciences
Discussions

Interactive comment on “Data compression to define information content of hydrological time series” *by* S. V. Weijs et al.

S. V. Weijs et al.

steven.weijs@epfl.ch

Received and published: 13 May 2013

We thank the reviewer for his/her thoughtful comments, questions and suggestions, which will be very helpful in improving the paper. Below we respond to the comments (*repeated in italic font*) point by point.

In this paper the authors try to introduce the AIT to hydrologists, and to present the results of application of same compression algorithms to hydrological data. They try to explain how these algorithms could reflects the information content of these data. Although the bibliographical review is needed to understand the paper because the AIT could be new for the readers of the journal (hydrologists), it seems to be too long. Some paragraphs of this part could be omitted for brevity.

We believe that Algorithmic Information Theory (AIT) is new to almost all readers of the journal/hydrologists in general, since a search for papers citing the key papers of Solomonoff, Kolmogorov and Chaitin gave zero results for papers published in HESS and the other major hydrology journals, except for one earlier paper by the first author that briefly pointed to it. AIT derives its power mainly from its deep foundations and less from its practical applicability; we therefore think an accessible short introduction to AIT with references to its foundations is essential to back up the underlying philosophy of the compression approach, which is an approximation to AIT to bridge the gap to practice. AIT itself can provide a quantitative and well-founded view on many issues currently discussed in hydrological literature, and the introduction in this paper is intended to inspire ideas based in this way of thinking for the hydrological community. We agree that there is a lot of background and introductory material, which we think is necessary to introduce the theoretical basis of AIT in the hydrological context. This is an important objective of the paper. Does the reviewer’s comment about length concern the entire section 2 or specific parts? We will try to shorten where possible, but the second reviewer also asked for more detailed explanations in this section, so we are afraid it will be difficult to shorten much without losing clarity and content.

The authors don’t show clearly the usefulness of the results obtained in hydrology, especially what is the supply of AIT compared to the data mining techniques used in hydrology.

We will clarify that AIT is a theory underlying inference problems and data mining techniques can be viewed as practical techniques which can be explained in terms of AIT as underlying theory. AIT gives the bounds on what is possible and impossible and gives insights in assumptions underlying commonly used techniques. Any practical technique for inference must make such assumptions to be computable and AIT could serve as a golden, but incomputable standard to make explicit what these assumptions are. We also will add some discussion on what can practically be done with estimations of information content obtained from compression experiments.

The paper is interesting, the methodology is clearly described; however, the most important results are left as a question for future researches. It seems also that the authors let the important results (application with hydrological model) to a future paper. It would be better that that the two papers (present and futur paper) were published in the same journal as part 1 and part2.

Indeed this paper should be seen as a first step, introducing AIT to hydrologists and showing how AIT as a theory reveals the inherent difficulties in defining information content, which result in limitations and subjectivity. The fact that these difficulties are inherent is often overlooked in the current debates. The results on single time series presented in this paper serve as illustration of AIT, as proof of concept of the compression approach and as a benchmark for a follow-up paper using a hydrology specific compressor to illustrate the dependence on prior knowledge. This last paper has just been published in the open access journal ”Entropy” and is therefore equally accessible as a part1 - part2 paper in the same journal would be, with the added advantage of reaching the broader readership of both journals. The future research about using hydrological models to jointly compress hydrological P and Q time series has not been completed yet and needs a complete new setup for which the algorithms still have to be developed. We hope to present this research, possibly in HESS, when it is finished and results have been produced and analyzed.

Quantization: the authors have chosen a uniform quantization with a precision equal to 8 bits. The question is why they have chosen 8 bits and how they have done this choice?

We will clarify this in the paper: Eight bits was chosen because most of the existing compression algorithms work at the byte (8 bits)-level. Using more than 8 bits per value would therefore not allow the algorithms to detect and utilize dependencies. This problem also applies to quantization using less than eight bits, but it is possible to simply quantize to eight bits but only use e.g. 6 of them, keeping the first 2 bits always at 0. This approach may be useful to illustrate the discussions on the dependence of

information content on the question asked (p2046:14-27; p2047:10-21) and the role of model complexity (p2048:3-18). We will make the link to this discussion clearer.

This choice could affect the results (by inducing a loss of precision). Indeed, one of the points this paper aims to illustrate is that information content is relative to the question that is asked (which is determined by the quantization chosen), and that this relativity is naturally revealed by the AIT framework.

It’s possible that the quantization smooth the data with large range (the case of rainfall data for humid region). Is it better to choose precision proportional to the range of data for example?

This smoothing is probably the case and this is one example of the sensitivity of information content to the question asked. The proportional scheme, in which the quantization has equal fixed width intervals for all time series will ask the data a question like ”in which 1 mm/day interval do to the precipitation values fall?”, instead of ”in which percentage block of the total range does the value fall?”. This quantization thus asks a different and equally defendable question, and the result will reflect the information content of the data relative to that question (still this information content ”subjectively” depends on prior knowledge about the answer to that question). The main problem with the quantization method we used is the sensitivity to a single peak value that determines the range. One could also chose a fixed quantization for all time series, probably making the results highly dependent on variance. This approach would ask another question to the data and one can argue about the relevance of that question, i.e. the drawbacks of the quantization scheme.

I recommend to do other experiments with different precision and compare results.

This would indeed be an interesting experiment which illustrates the subjectivity discussed in this paper. In fact, including the effect of quantization in the comparison changes the focus from lossless compression algorithms (which take the quantized series as the input to reproduce losslessly), to lossy compression, in which the trade-off between information loss (precision) and file size is considered. This can be interesting if related to measurement precision, but it is outside the scope of the present paper to discuss this in detail. We will add a short paragraph discussing lossless vs. lossy compression in this context.

It's possible also to use a non-uniform quantization (μ-law quantization for example) that could gives less loss of information.

Yes, this is certainly possible. Actually, we used a non-uniform quantization for Q, by first log-transforming it before quantizing with the uniform scheme. This is actually very similar Mu-Law quantization for the untransformed values. For discharge, the fact that the entropy of the quantized values is high (close to the maximum of 8 bits), indicates that the quantized signal contains a high amount of information compared to what is possible with 8 bits.

Assessing the information loss of the quantized compared to the original signal, is plagued by many problems of subjectivity again, since also the information content of the original signal depends on the question asked. Moreover, the question of data quality and information content of signal vs. noise come into play, since the original values are typically stored with much higher precision than the data quality would warrant. We will adapt the text to reflect these considerations.

In experiment B, why you use the same quantizatiion schema of Q for Qer, Why you don't use the limits (min and max) of Qer for quantization.

We will explain in the revised paper that we keep the same quantization to compare the information content of Q with and without knowing P and the hydrological model. Changing the quantization makes the results incomparable, because in that case the question we ask the data is changed when introducing the model (by changing quantization), and information content is relative to that changed question. When keeping the quantization the same, we can assess how much of the uncertainty in Q is explained by P.

Section 3.2 : I recommend to authors to give more references about the compression algorithms used and the references of their source code.

We will add more references to make the results better reproducible. We will present the necessary details (exact commands, links to software) to reproduce the results in the supplementary material.

Section 3.3 It seems better to explicit the different method used to generate the data. Readers could not understand the meaning of sine1, sin100..etc in tab 2.

The meanings are defined in tab 1, but indeed these may be hard to interpret unambiguously. We will improve the description and present the necessary details such as Matlab code in the supplementary material, so the signals can be exactly reproduced.

Section 5 : discussion and conclusion need to be more related to the experiment done by the authors.

Indeed the discussion and conclusions are mostly about the problems relating to the interpretation of compression results, which are inherent in the question of information content, independent of which method is used to assess it. We think it is essential that these limitations be understood, instead of devoting much discussion on interpretation of the results proper. To relate the discussion and conclusions more to the results, we will make the above clearer, possibly by discussing results of new experiments with different quantization. Furthermore, we will add some discussion on what could be done with the results for information content, assuming we have verified that the experiment done correctly reflects our question and prior knowledge. Finally, we will also make clear in the conclusions how the experiment done on single time series relates to future planned experiments on joint compression of rainfall runoff data.

in page 2033 line 17 please say that the DKL is the KullBack-Leibel divergence and give a reference.

We now added the full form and reference for the Kullback-Leibler divergence.

Interactive comment on Hydrol. Earth Syst. Sci. Discuss., 10, 2029, 2013.