Hydrol. Earth Syst. Sci. Discuss., 10, C851–C854, 2013
www.hydrol-earth-syst-sci-discuss.net/10/C851/2013/

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.

Anonymous Referee #2

Received and published: 16 April 2013

General remarks: To derive the information content of hydrologic series, the article explores the potential of a number of data compression methods which are broadly adopted in computer sciences, signal and image processing as well as telecommunications to reduce data storage and transmission capacities. In computer science and signal processing the linkage with information theory lies in the fact that data compression (and decompression) needs to be achieved with a minimum loss of information. In hydrology, information theory has been so far adopted using entropy concept in hydrological frequency analysis or to size hydrological networks (decision problems) or to adjust rainfall-runoff models using specific cost functions and more recently as a way to define the information content of data in regionalization or sampling problems. Therefore, the linkage between data compression technology and hydrological regionalization which is the main focus of the paper is really a very interesting and vast

domain of investigation in hydrology. However, it is not very frequent to find entropy expressed in bits neither in hydrological frequency analysis nor in network optimisation problems. So, I found that the style of paper and its writing do not help readers from the hydrological and climatological fields to understand the potential of the connexion with computer sciences and signal processing disciplines. Therefore, my general recommendation would be to "smooth" the text in order to make it easier for readers from the hydrology side (as supposed are HESS readers) and to give more details about the hydrological context. In order to help interpret the information content obtained it seems important to give an idea about hydrological data at least for the Leaf river basin (some basin physiography characteristics, sample statistics and main hydrological signatures). For basins from MOPEX project, we would need to give an idea about basin climate classification, area, average annual rainfall and runoff etc. . . which represent the hydrological context of the study. Otherwise, this article would be more suitable in an informatics journal, addressed to capture informatics scientists about the potential of their methods in the hydrological field.

Specific remarks (P 4) The introduction of 2.1 is difficult (lines 8 to 14). A definition of the code function seems to be necessary before introducing code lengths. Also, the definition of the prefix code would be helpful at the beginning of the paragraph. The binary tree of Fig. 1 should be well introduced (the size, the height, the nodes, the roots, the leaves, branches . . .) in linkage with data structure and not only on the basis probability. The term event should be defined in relation with the tree representation as well as the length of the code (tree height, depth (distance to root node), leaves, nodes). It is not relevant to put such details in the appendix. Line 20 "see fig 1 code A" is not enough explicit of what you mean; you should explain this example in more details. Huffman coding needs to be better introduced and documented as well as Range coding (P16 line 4). The Kraft inequality (Eq. 1) and Eq. 3 (Kraft- MacMillan theorem) may be presented in a more comprehensive way. In particular you should give the definition of the Kullback-Leibler divergence in simple words ("a measure of the information lost when q is used to approximate p") in this part of the text. P5 line 2

the terminology of “bit per sample” might be specified here. I think that the Appendix is not enough for the reader who is not used with this terminology or who is interested to go further in its application for his own purposes. P5 in eq(2) you might specify the base of the logarithm (base 2 because you use the bits units?) P5 line 10 the equation you are referring to is not specified (Eq. 3).

P11 EQ. 5 minx , maxx and xinteger are not specified P11 line 15 “The compression algorithms will be mainly used to explore the difference in information content between different signals “. How did you explore this idea in the results analysis? P13 line 20 the generation procedure should be described shortly (sinus etc. . .?) Line 22 14610 potential evapotranspiration (it is not potential evaporation) P14 l 12-14: the text is not clear P15 line 7 the byte definition might be recalled otherwise the understanding of number 256 in “256 unit values” would not be direct. P15 line 10 “by value” has to be removed P16 line 10 I could not understand your findings. Low entropy results in high predictability. In your case, are streamflow series more predictable or less predictable than precipitation series?; In Table 2 H/log N for LEAFQ (=42.1)>H/log N for LEAFP (=31) indicating that streamflow series are more unpredictable than rainfall series. It is also the case in Fig. 3 for Mopex watersheds. How would you explain these results while the autocorrelation in rainfall series (=0.15) is far less than in runoff series (=0.89) (Table 3)? What kind of daily time series did you adopt for precipitations? Did you use spatial average daily rainfall for a given watershed or a single raingage in the outlet of the watershed or inside the watershed? Did you control the fact that precipitation and discharge data are well compatible (by calculating runoff coefficients for example)? For the lossless compression context as stated in p9 line 12., the same quantity of information as the original is carried out using fewer bits; this leads to more information per bit which is equivalent to more entropy. Does Fig. 3 represent the information per bit ? On the other hand, the term better compressible should be explained (comparison of bit of information per bit of message or compression size normalized by entropy?) . The better compressibility of streamflow data should be interpreted in this part of the text (such as to be linked to watershed size and basin geological features) (Fig. 4a)

P17 line 11 what is the purpose of the study of errors compression? Rainfall-runoff discharge errors are generally autocorrelated. What does it indicate relatively to model structure or performances? Lower entropy of the errors means lower unpredictability of errors. Here you are right to mention that interpretation of entropy in terms of data compression is not simple; P17 line 15 too complex

---

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