Geosci. Model Dev. Discuss.,
https://doi.org/10.5194/gmd-2018-250-RC1, 2019

Geoscientific
Model Development
Discussions

Interactive comment on "Evaluation of lossless and lossy algorithms for the compression of scientific datasets in NetCDF-4 or HDF5 formatted files" *by* Xavier Delaunay

Anonymous Referee #1

Received and published: 16 January 2019

Manuscript: "Evaluation of lossless and lossy algorithms for the compression of scientific datasets in NetCDF-4 or HDF5 formatted files"

Authors: Xavier Delaunay, Aurélie Courtois, and Flavien Gouillon

——————————- Summary ——————————-

The authors focus on data compression (lossless and lossy) of HDF5 or NetCDF4 files. They compare several techniques on synthetic data and mission data and also suggest an alternative lossy method called Digit Rounding (to improve upon Bit Grooming). Data compression of scientific data is an important topic that is continuing to receive attention. This manuscript investigates compression on some synthetic data and

application data, and draws general conclusions based on those studies.

——————————- General comments ——————————-

- This manuscript needs a lot of improvement in terms of the grammar and writing. There are many awkward phrases and incorrect word choices that need to be improved (a subset are listed below). The paragraph structures are also in need of modification (many paragraphs contain only 1 or 2 sentences).

-Section 2: I'd be helpful to include more detail for the preprocessing algorithms: shuffle and bitshuffle. Also this section in general needs improvement. It's a bit "choppy" to read (needs smoother and better transitions between topics) and feels like more details would be helpful on the methods (especially the ones that the digit rounding algorithm builds on).

-Section 4: Why does using the synthetic data in 4.1 to assess performance make sense - it seems unrelated to the application area of interest. I'd argue that the metrics used in 4.1 are really minimal requirements as well. Also take care when referring to "performance" as it is overloaded term...do you mean speed or effectiveness (it's used both ways)

-fpzip is a fast and effective lossless method that would have been nice to compare (I *think* there is an fpzip filter available). Also I believe that any hdf5 filter can be accessed through NetCDF4 (see last sentence in conclusion) - consider contacting the Unidata folks.

-Comments on doing compression in parallel?

-When reading the conclusion, it's hard to see what the main contributions of this paper are. It's fairly well known already that preprocessing of scientific data (e.g., bit shuffle or shuffle) improves lossy compression. Also the statements in the conclusion aren't specific to a particular type of data set, but are presented as more general conclusions. Given that the effectiveness and performance (speed) of lossy and lossless compression are very data, application, and variable dependent, the general statements here are not well justified by the small sample of data in the paper. I'd suggest focusing the paper more heavily on the data in Section 5 (if it's of interest) and tailoring the discussion in that manner. Or maybe the focus was to be more on speeds than quality, in which case it's be important to work to get sz and fpzip working, particularly via netcdf-4...

——————————- Specific items: ——————————-

-p2, line 20: note that fpzip can also be lossless

-p. 8: discussion of figure 5: is the width of the bars related to the compression levels? (e.g. line 20 statement is unclear)

-p.8, lines 28-29: Why is this the case? (Add some discussion beyond describing the figure.)

-p.8, line 20: I feel like the parameters should be better explained for –filter so that the reader can try them more easily. For example, what does the "32017" mean? I think that the following 0 is for sz, but this is not stated either.

-p. 10, line 6: re: "experiments have shown" - whose or which experiments (cite?)

-Table 5 : Why is the speed faster for 1d?

-Section 4.4.1, line 21-24: Any idea why you get these results?

-Section 4.4.1, last sentences: It's unclear to me what the value of these synthetic data sets is - especially given the statement on p. 11, line 3, about the dependence on the dataset

-page 11, line 9-10: I'd include characteristics of the data (e.g., maximum abs. value) earlier in the text when the two datasets are introduced.

-page 11, line 24: I don't see relative error mentioned in Table 6 - it seems to just be

absolute error

-p.11-12: Need more of a discussion of the results in Figure 7. For 3D, it looks like bit grooming and digit rounding are similar - I don't see a clear advantage.

-p.12, lines 16-17: SZ compression can be controlled with an absolute error bound, so why is the relative error bound adjusted to get the desired abs. error?

-Section 5.1: It is disappointing not to have SZ results on the real data of interest. Were the SZ authors contacted? I would think that they could have helped resolve this issue.

-p. 15, line 18: "which only a few attributes may be missing" - It's unclear what this means. It's super helpful to really detail the data being compressed so that one can make sense of the results.

-p. 14, line 30: Please share more specific information about the precision required by the scientists for the data. Again, more information is useful for interpreting results.

-Section 5 seems like it should be the highlight of the paper as here we are seeing the results on the real data. But it feels like more detail is needed on the data and more discussion of the implications of the results.

————————— Typos, etc.: —————————

-abstract, line 7: incorrect use of "imposes"

-p.1, line 22: "quite spread"=> "quite prevalent" or "quite popular" , "widely spread" => "widely used"

-p.1., line 26: "reduce significantly" => "significantly reduce"

-p.1. line 27: This sentence (that continues to page 2) is too long.

-p.2, line 3: "can afford for" is awkward

-p2, lines 10-24: this region is 5 paragraphs

-p.3, lines 3-4: awkwardly worded

-p. 3, line 9: one sentence paragraph

-p.3, line 12: missing "," after "Deflate"

-p.3 line 14: not sure what is meant by "new concurrent"

-p.3, line 13: "widely spread" => "widely used"

-p.3, line 16: awkward sentence: "This allows Deflate achieving rather high compression ratios"

-P.3, section 3: again, there are too many tiny paragraphs

-p.4, line 7: "are of same interest" is awkward

-p.4, line 21: Table number is not given

-p.5, line 15: awkwardly worded

-p.5, line 24: One sentence paragraph

-p.7, section 4.2: define f_s, f_ech

-p.7, line 22: "use embarks" is awkward

-p.8, line 7: "declined" doesn't make sense

-p.8, line 14: "embark" - incorrect usage

-p.10, line 10: "This correspond corresponding" needs to be fixed

-p. 10, line 24: Note sure I'd use "performances" here as earlier it was used to indicate speed.

-p. 15, line 21: "ration" => "ratio"

-p.15, line 14: another one sentence paragraph

-p.16, line 9: "Extends to this work" - awkwardly worded

Interactive comment on Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2018-250, 2018.