Geosci. Model Dev. Discuss.,
https://doi.org/10.5194/gmd-2018-250-RC2, 2019

Geoscientific Model Development Discussions
Open Access
EGU

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

Zender (Referee)

zender@uci.edu

Received and published: 26 January 2019

I have voluntarily disclosed my identity in all manuscript reviews since 2004. The authors are free to contact me at zender@uci.edu.

General Comments

This manuscript presents a new lossy compression algorithm called "Digit Rounding" (DR), and evaluates its performance against and with other lossy and lossless compression algorithms on idealized and remote sensing datasets. The manuscript addresses the growing need to archive meaningful data rather than noise, and to do

Printer-friendly version

Discussion paper

so reliably and quickly. The study presents an original advance in lossy compression whose implementation unfortunately hampers its utility. The study is understandable yet poorly written. This potentially useful study of lossy compression techniques needs a thorough overhaul before publication.

Specific Comments

Originality: DR is an improvement on “Bit Grooming” (BG) which I invented as an improvement on “Bit Shaving”. In that sense I am qualified to comment on its originality. The heart of DR is essentially a continuous version of BG: Whereas BG fixes the number of bits masked for each specified precision, and masks these bits for every value, DR recomputes the number of bits masked for each quantized value to achieve the same precision. BG did not implement the continuous method because I thought that computing the logarithm of each value would be expensive, inelegant, and yield only marginally more compression. However, DR cleverly uses the exponent field instead of computing logarithms, and so deciphers the correct number of bits to mask while avoiding expensive floating point math. This results in significantly more compressibility that (apparently) incurs no significant speed penalty (possibly because it compresses better and thus the lossless step is faster?). Hence DR appears to be a significant algorithmic advance and I congratulate the authors for their insight.

The manuscript stumbles in places due to low quality English, and cries out for more fluent editing. Not only is the word choice often awkward, but the manuscript is like a continuously choppy sea of standalone sentences with few well developed paragraphs that swell with meaning then yield gently to the next idea. GMD readers deserve and expect better.

Does DR guarantee that it will never create a relative error greater than half the value of the least significant digit? BG chooses the number of digits to mask conservatively, so it can and does guarantee that it *always* preserves the specified precision. Equations

(1)-(7) imply that DR can make the same claim, but this claim is never explicitly tested or made. The absence of this guarantee is puzzling because it would strengthen the confidence of users in the algorithm. However, the guarantee must be explicitly tested, because it undergirds the premise that the comparison between DR and BG is fair. In any case, clearly state whether DR ever violates the desired precision, even if that happens only rarely.

p. 16 L13: "Code and data availability: The Digit Rounding software source code and the data are currently only available upon request to Xavier Delaunay (xavier.delaunay@thalesgroup.com) or to Flavien Gouillon (Flavien.Gouillon@cnes.fr)." The GMD policy on code and data is here: https://www.geoscientific-model-development.net/about/code_and_data_policy.html. This manuscript provides no code access nor explanation, and no dataset access, and thus appears to violate GMD policy in these areas.

Common comparisons would help build confidence in your results. It would have been more synergistic to evaluate the algorithms on at least one of the same datasets as Zender (2016), which are all publicly available. I am glad the authors used the publicly available NCO executables. Why not release the DR software in the same spirit so that the geoscience community can use (and possibly improve) it?

The lossless and lossy compression algorithms analyzed seem like a fairly balanced collection of those most relevant to GMD readers. Most methods that were omitted are, to my knowledge, either non-competitive (e.g., Packing) or not user-friendly, e.g., research grade but not widely available (e.g., Layer Packing) and too hard to independently implement.

Table 6 on p. 19 shows the maximum absolute error (MAE) of BG is quite similar to DR, as I would expect. However, Table 7 on p. 20 shows the maximum absolute error (MAE) of BG is nearly 10x less than DR. Why are the MAEs similar for dataset s1

and significantly different for dataset s3D? I expect DR has a greater mean error (and lower SNR) than BG due to the algorithms, yet the difference in MAEs surprises me. Zender (2016) Table 3 shows that BG is tuned to have an MAE just shy of violating the precision guarantee. An MAE that is nearly 10x larger seems like it might violate the precision guarantee.

The preceding comment is a request to more carefully analyze the underlying cause of the behaviors reported in the data. The next two comments are to report more results to deepen the analyses and explain the behavior of DR more robustly.

Please include the maximum absolute error or maximum absolute relative error (which normalizes the error by the original value) to Tables 5–10.

MeanAE is an important statistic that is complementary to MaxAE. MeanAE is the average absolute (no compensation between positive and negative) bias in the dataset, and is more familiar and relevant than SNR to at least some geophysicists. Please consider including MeanAE in Tables 5–10.

Zender (2016) and Silver and Zender (2017) consider four primary criteria to evaluate compression algorithms: Compression Ratio, Accuracy, Speed, and User-friendliness. This manuscript neglects explicit consideration of the last, though usability seems (in addition to performance) seems to be an implicit reason why they recommend BG not DR for the “real world” use cases in Sections 5.1 and 5.2. The manuscript would benefit from a more explicit consideration of usability throughout. Examples include software availability, flexibility, and complexity of invocation, as well as transparency (will users have all the necessary software required to read the compressed data?), and instructions to mitigate these issues for DR.

Tables 1 and 3 follow Tables 1 and 2 of Zender (2016). This should be noted in the text and/or caption of the tables.

It seems like Table 2, the algorithm description, should be a figure rather than a table.

The manuscript is awkward in that it introduces a demonstrably superior lossy compression algorithm but recommends a different algorithm (BG) for "real world" cases (Section 5), partly because DR is unavailable in software that potential users have easy access to, and its implementation appears to be too inflexible to use on generic datasets. The recommendation of BG not DR does attest to the objectivity of the study, yet it seems to be an unsatisfying conclusion to what was clearly a time-consuming study. In this sense the manuscript seems premature, since if DR were "ready for primetime" then the authors could have recommended it rather than BG in Section 5. Perhaps the authors should re-evaluate whether the manuscript is premature, i.e., whether it should both introduce a new lossy algorithm before it is ready to use in optimized workflows for generic geoscientific data compresssion.

Minor Suggestions

p. 1 L22: "well spread"

p. 2 L22: DEFLATE

p. 4 L1: $\max_i$ is redundant. Just use $\max$.

p. 4 L21: Table 1

p. 9 L7: "declined"?

p. 9 L14: "By default, Sz algorithm embark Deflate." is awkward.

p. 14 L27–28: These lines are identical

p. 18 L8: "the number d_i of significant digit number of digits"???

p. 18 L8: “following Eq.” not “following in Eq.”

p. 23 Figure 4: Clarify the meaning of the distinct vertical bars.

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

Printer-friendly version

Discussion paper