Geosci. Model Dev. Discuss.,
doi:10.5194/gmd-2016-63-AC1, 2016

Geoscientific Model Development Discussions

Interactive comment on "Bit Grooming: Statistically accurate precision-preserving quantization with compression, evaluated in the netCDF Operators (NCO, v4.4.8+)" *by* Charles S. Zender

C. S. Zender

zender@uci.edu

Received and published: 31 May 2016

(Reviewer's comments/questions are in *italics* and my responses are interspersed in plain text.)

I thank the Reviewer for their thoughtful comments. I share the Reviewer's perspective that these techniques are underappreciated in the Geosciences, and am glad to help rectify that in a small way.

In the Abstract on line 13, it is mentioned that Bit Grooming produces storage reductions comparable to other quantization techniques such as linear packing when "used

aggressively". Is this always true?

The wording of this question makes it important to clarify for others that the manuscript asserts that it is Bit Grooming (not, e.g., Packing) that must be used aggressively to match compression ratios (CRs) produced by other techniques (e.g., Packing). Standard Packing (float32->short16) of float-dominated data always produces CR of about 50%. This CR is intrinsic to the Packing algorithm and applies to any float-dominated dataset (the only type of dataset this manuscript discusses).

Our results show lossless compression further reduces Packing-assisted CRs to about 20% (relative to uncompressed data). In every case tested Bit Grooming must be used aggressively (i.e., preserve at most 2 significant digits) to match or best these CRs. I chose the test data to be representative, and know of no real-world float-dominated datasets where Bit Grooming CRs could match or best Packing CRs unless Bit Grooming were used this aggressively. So the answer to your original question, as I interpret it, is "Yes". The manuscript now clarifies that "aggressively" means preserving only 1 or 2 digits.

On line 22, the statement that begins "False precision can mislead..." and the following sentences express a concept that should be captured in the abstract. This is the real strength of this approach: turning useless precision into something that is (a) more honest, and (b) saves space!

Agreed. Unless a reviewer objects, the abstract in the revised manuscript will include roughly the same content as the three sentences you refer to. The working draft abstract now has a longer first paragraph (second paragraph is unchanged):

"Geoscientific models and measurements generate false precision (scientifically meaningless data bits) that wastes storage space. False precision can mislead (i.e., imply noise is signal) and be scientifically pointless, especially for measurements. By contrast, lossy compression can be both economical (save space) and heuristic (clarify data limitations) without compromising the scientific integrity of data. Data quantiza-

tion can thus be appropriate regardless of whether space limitations are a concern. We introduce, implement, and characterize a new lossy compression scheme suitable for IEEE floating-point data. Our new Bit Grooming algorithm alternately shaves (to zero) and sets (to one) the least significant bits of consecutive values to preserve a desired precision. This is a symmetric, two-sided variant of an algorithm sometimes called Bit Shaving which quantizes values solely by zeroing bits. Our variation eliminates the artificial low-bias produced by always zeroing bits, and makes Bit Grooming more suitable for arrays and multi-dimensional fields whose mean statistics are important."

The "eight-hundred pound gorilla" example is cute, but perhaps a better example would be something less cute and ordinary, such as a "liter of milk" or something.

It is important that the example have more than one digit, and also that some digits be insignificant, i.e., that the quantity be recognized as an approximation that is not exact. And finally the example must be dimensional and denominated in a standard unit like mass, volume, or time. A "liter of milk" won't work, neither will 10 or 100 liters because milk bottles are measured in exact units with high precision. I don't see the drawback of the gorilla example, which has the necessary properties. Ordinary examples can be good, and cute examples can increase readers' interest and retention.

It's great that the source code is provided on Github. Kudos to the authors for making the code truly open source!

Thank you for appreciating the importance of this!

Interactive comment on Geosci. Model Dev. Discuss., doi:10.5194/gmd-2016-63, 2016.